Uniform convergence over time of a nested particle filtering scheme for recursive parameter estimation in state–space Markov models

Dan Crisan* Joaquín Míguez†

March 31, 2016

Abstract

We analyse the performance of a recursive Monte Carlo method for the Bayesian estimation of the static parameters of a discrete–time state–space Markov model. The algorithm employs two layers of particle filters to approximate the posterior probability distribution of the model parameters. In particular, the first layer yields an empirical distribution of samples on the parameter space, while the filters in the second layer are auxiliary devices to approximate the (analytically intractable) likelihood of the parameters. This approach relates the this algorithm to the recent sequential Monte Carlo square (SMC^2) method, which provides a *non-recursive* solution to the same problem. In this paper, we investigate the approximation, via the proposed scheme, of integrals of real bounded functions with respect to the posterior distribution of the system parameters. Under assumptions related to the compactness of the parameter support and the stability and continuity of the sequence of posterior distributions for the state–space model, we prove that the L_p norms of the approximation errors vanish asymptotically (as the number of Monte Carlo samples generated by the algorithm increases) and uniformly over time. We also prove that, under the same assumptions, the proposed scheme can asymptotically identify the parameter values for a class of models. We conclude the paper with a numerical example that illustrates the uniform convergence results by exploring the accuracy and stability of the proposed algorithm operating with long sequences of observations.

*Department of Mathematics, Imperial College London (UK). E-mail: d.crisan@imperial.ac.uk.

†School of Mathematical Sciences, Queen Mary University of London (UK). E-mail: j.miguez@qmul.ac.uk.

1 Introduction

The problem of parameter estimation arises in a multitude of applications of state–space dynamic models and, as a consequence, has received considerable attention from different perspectives [20, 24, 1, 17, 4, 18]. We investigate the use of a nested particle filtering scheme, introduced in [10], for the recursive Bayesian estimation of the static parameters of discrete-time state-space Markov systems.

1.1 Background

To ease the presentation, let us consider two (possibly vector-valued) random sequences $\{X_t\}_{t=0,1,\ldots}$ and $\{Y_t\}_{t=1,2,\ldots}$ representing the (hidden) state of a dynamic system and some related observations, respectively, with t denoting discrete time. The state process is assumed to be Markov and the observation Y_t is independent of any other observations $\{Y_k; k \neq t\}$, conditional on the state X_t. The conditional probability distribution of X_t given $X_{t-1} = x_{t-1}$ and the probability density function (pdf) of Y_t given $X_t = x_t$ are assumed to be known up to a vector of static random parameters, denoted Θ. These assumptions are very common in the literature and actually hold for many practical systems (see, e.g., [31, 3]). Given a sequence of observations, $Y_1 = y_1, \ldots, Y_t = y_t, \ldots$, the Bayesian parameter estimation problem consists in tracking the posterior probability distribution of the parameter vector Θ over time.

When the parameter vector is known, $\Theta = \theta$, it is a common approach to use particle filters [16, 19, 25, 15, 29, 14, 31, 3, 22] in order to track (over time t) the posterior probability distribution of the state X_t conditional the record of observations, $Y_{1:t} = y_{1:t}$, which is often termed the *filtering distribution*. At each time step, a particle filter generates a discrete random approximation of the filtering distribution that consists of samples on the state space. Unfortunately, the design of particle filtering methods that can account for a random vector of parameters in the dynamic model (i.e., a static but unknown Θ) is a hard problem and it has remained an open issue for two decades. While many algorithms have been proposed [23, 5, 24, 32, 1, 28, 4, 30] none of them is widely accepted as a complete solution to this problem. Some of them are seen as *ad hoc* [24], others depend on the structure of the state–space model to be applicable [5, 32, 4] and others yield only point estimates rather than approximations of the sequence of posterior distributions [23, 1, 30]. The recent sequential Monte Carlo square (SMC^2) method [6] overcomes these problems, but the algorithm is *not* recursive and hence it becomes computationally prohibitive when the sequence of observations is relatively long. See [18] for a recent survey of the field.

1.2 Contributions

We investigate the convergence and performance of the nested particle filtering scheme in [10] for the approximation of the posterior distribution of the unknown parameters Θ given the data $Y_{1:t} = y_{1:t}$. Similar to [28] and [6], the algorithm consists of two nested layers of particle filters: an "outer" filter that approximates the probability measure of Θ given the observations and a bank of "inner" filters that yield

approximations of the posterior probability distribution of X_t conditional on specific realisations of Θ. The outer filter directly provides an approximation of the marginal posterior distribution of Θ, which is the main object of interest in this paper. The proposed scheme is similar to the SMC^2 method of [6]. However, unlike SMC^2, it is a purely recursive procedure that readily admits an online implementation. A detailed comparison of the two algorithms is provided in [10].

In this paper we look into the approximation, via the proposed scheme, of integrals of real bounded functions with respect to (w.r.t.) the posterior distribution of the system parameters. Under a set of assumptions related to

- the compactness of the parameter space,
- the stability of the sequence of posterior probability measures associated to Θ and X_t, and
- the continuity of the conditional (on Θ) optimal filters in the state–space

we prove that the L_p norms of the approximation errors vanish asymptotically, as the number of particles in the filter increases, and uniformly over time. In particular, we obtain an explicit upper bound for the L_p approximation errors that is independent of the time index t. This uniform convergence result has some relevant consequences. One of them is that the proposed scheme can eventually identify the parameter values for a broad class of state-space models. In particular, we prove that, when the true posterior probability measure of Θ converges toward a unit delta measure located at a point θ_* in the parameter space, the approximation computed via the proposed nested particle filter also converges to the same delta, in terms of a suitable distance, as $t \to \infty$.

In order to illustrate the theoretical results, we present computer simulation results, for a stochastic Lorenz 63 model, which show numerically how the nested particle filtering algorithm attains an accurate and stable performance with a fixed number of particles and long sequences of observations.

1.3 Organisation of the paper

We present a general description of the random state-space Markov models of interest in this paper in Section 2. In Section 3 we describe the proposed nested particle filtering scheme. A summary of the theoretical findings in the paper is provided in Section 4, while the full analysis of the algorithm is described in Section 5. In Section 6 we present the results of our computer simulation experiments. Finally, Section 7 is devoted to the conclusions.

2 Background

2.1 Notation, assumptions and preliminary results

We first introduce some common notations to be used through the paper, broadly classified by topics. Below, $\mathbb{R}$ denotes the real line, while for an integer $d \geq 1$, $\mathbb{R}^d = \overbrace{\mathbb{R} \times \ldots \times \mathbb{R}}^{d \text{ times}}$

- Functions: Let $S \subseteq \mathbb{R}^d$ be a subset of $\mathbb{R}^d$.
  - The supremum norm of a real function $f : S \to \mathbb{R}$ is denoted as $\|f\|_\infty = \sup_{x \in S} |f(x)|$.
  - $B(S)$ is the set of bounded real functions over S, i.e., $f \in B(S)$ if, and only if, $\|f\|_\infty < \infty$.
  - We use $a \vee b$ and $a \wedge b$ to denote the maximum and the minimum, respectively, between two real numbers a and b.
- Measures and integrals:
  - $\mathcal{B}(S)$ is the σ-algebra of Borel subsets of S.
  - $\mathcal{P}(S)$ is the set of probability measures over the measurable space $(S, \mathcal{B}(S))$.
  - $(f, \mu) \triangleq \int f(x)\mu(dx)$ is the integral of a real function $f : S \to \mathbb{R}$ w.r.t. a measure $\mu \in \mathcal{P}(S)$.
  - Given a probability measure $\mu \in \mathcal{P}(S)$, a Borel set $A \in \mathcal{B}(S)$ and the indicator function
$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases},$$
$\mu(A) = (I_A, \mu) = \int I_A(x)\mu(dx)$ is the probability of A.
- Sequences, vectors and random variables (r.v.'s):
  - We use a subscript notation for sequences, namely $x_{t_1:t_2} \triangleq \{x_{t_1}, \ldots, x_{t_2}\}$.
  - For an element $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, its Euclidean norm is denoted as $\|x\| = \sqrt{x_1^2 + \ldots + x_d^2}$.
  - The L_p norm of a real r.v. Z, with $p \geq 1$, is written as $\|Z\|_p \triangleq E[|Z|^p]^{1/p}$, where $E[\cdot]$ denotes expectation w.r.t. the probability distribution of Z.

Remark 1 *Let $\alpha, \beta, \bar{\alpha}, \bar{\beta} \in \mathcal{P}(S)$ be probability measures and let $f, h \in B(S)$ be two real bounded functions on S such that $(h, \bar{\alpha}) > 0$ and $(h, \bar{\beta}) > 0$. If the identities*

$$(f, \alpha) = \frac{(fh, \bar{\alpha})}{(h, \bar{\alpha})} \quad \textit{and} \quad (f, \beta) = \frac{(fh, \bar{\beta})}{(h, \bar{\beta})}$$

hold, then it is straightforward to show (see, e.g., [8]) that

$$|(f, \alpha) - (f, \beta)| \leq \frac{1}{(h, \bar{\alpha})} \left|(fh, \bar{\alpha}) - (fh, \bar{\beta})\right| + \frac{\|f\|_\infty}{(h, \bar{\alpha})} \left|(h, \bar{\alpha}) - (h, \bar{\beta})\right|. \tag{1}$$

2.2 State-space Markov models in discrete time

Consider two random sequences, $\{X_t \in \mathcal{X}\}_{t \geq 0}$ and $\{Y_t \in \mathbb{R}^{d_y}\}_{t \geq 1}$, and a random variable $\Theta \in D_\theta$, where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $D_\theta \subset \mathbb{R}^{d_\theta}$ and the positive integers d_x, d_y and d_θ determine the dimension of the state space, the observation space and the parameter space, respectively. We further assume that D_θ is compact. Let $\mathbb{P}_t$ be the joint probability measure for the triple $(\{X_n\}_{n \leq t}, \{Y_n\}_{1 \leq n \leq t}, \Theta)$, that we assume to be absolutely continuous w.r.t. the Lebesgue measure.

The sequence $\{X_t\}_{t\geq 0}$ is the state (or signal) process, a possibly inhomogeneous Markov chain governed by an initial probability measure $\tau_0 \in \mathcal{P}(\mathcal{X})$ and a sequence of transition kernels $\tau_{t,\theta} : \mathcal{B}(\mathcal{X}) \times \mathcal{X} \to [0,1]$ indexed by a realisation of the r.v. $\Theta = \theta$. To be specific, we define

$$\tau_0(A) \triangleq \mathbb{P}_0\{X_0 \in A\}, \tag{2}$$

$$\tau_{t,\theta}(A|x_{t-1}) \triangleq \mathbb{P}_t\{X_t \in A|X_{t-1} = x_{t-1}, \Theta = \theta\}, \quad t \geq 1, \tag{3}$$

where $A \in \mathcal{X}$ is a Borel set. The sequence $\{Y_t\}_{t\geq 1}$ is termed the observation process. Each r.v. Y_t is assumed to be conditionally independent of other observations given X_t and Θ, namely

$$\mathbb{P}_t\{Y_t \in A|X_{0:t} = x_{0:t}, \Theta = \theta, \{Y_k = y_k\}_{k\neq t}\} = \mathbb{P}_t\{Y_t \in A|X_t = x_t, \Theta = \theta\}$$

for any $A \in \mathcal{B}(\mathbb{R}^{d_y})$. Additionally, we assume that, for every $x \in \mathcal{X}$ and $\theta \in D_\theta$, the r.v. $Y_t|X_t = x, \Theta = \theta$ has an associated probability density function (pdf). In particular, for some possibly unknown normalisation constant c, there are functions $g_{t,\theta}(y|x)$ such that

$$\mathbb{P}_t\{Y_t \in A|X_t = x_t, \Theta = \theta\} = c\int I_A(y)g_{t,\theta}(y|x_t)dy.$$

We assume that c is independent of y, x and θ.

If $\Theta = \theta$ (the parameter is given), then the stochastic filtering problem consists in the computation of the posterior probability measure of the state X_t given the parameter and a sequence of observations up to time t. Specifically, for a given observation record $\{y_t\}_{t\geq 1}$, we seek the measures

$$\phi_{t,\theta}(A) \triangleq \mathbb{P}_t\{X_t \in A|Y_{1:t} = y_{1:t}, \Theta = \theta\}, \quad t = 0, 1, 2, ...$$

where $A \in \mathcal{X}$. For many practical applications, the interest actually lies in the computation of integrals of the form $(f, \phi_{t,\theta})$ for some integrable function $f : \mathcal{X} \to \mathbb{R}$. Note that, for $t = 0$, we recover the prior signal measure, i.e., $\phi_{0,\theta} = \tau_0$ independently of θ.

We also introduce the predictive measure

$$\xi_{t,\theta}(A) \triangleq \mathbb{P}_t\{X_t \in A|Y_{1:t-1} = y_{1:t-1}, \Theta = \theta\}, \quad t = 0, 1, 2, ...,$$

which is closely related to the filter $\phi_{t,\theta}$ and we often write as $\xi_{t,\theta} = \tau_{t,\theta}\phi_{t-1,\theta}$, meaning that, for any integrable function $f : \mathcal{X} \to \mathbb{R}$, we obtain

$$(f, \xi_{t,\theta}) = \int\int f(x)\tau_{t,\theta}(dx|x')\phi_{t-1,\theta}(dx') = ((f, \tau_{t,\theta}), \phi_{t-1,\theta}). \tag{4}$$

Let us note that $\int f(x)\tau_{t,\theta}(dx|x')$ is itself a map $\mathcal{X} \to \mathbb{R}$. Integrals w.r.t. the filter measure $\phi_{t,\theta}$ can be rewritten by way of $\xi_{t,\theta}$ as

$$(f, \phi_{t,\theta}) = \frac{(fg_{t,\theta}^{y_t}, \xi_{t,\theta})}{(g_{t,\theta}^{y_t}, \xi_{t,\theta})}, \tag{5}$$

where $g_{t,\theta}^{y_t}(x) \triangleq g_{t,\theta}(y_t|x)$ is the likelihood of x. Eqs. (4) and (5) are used extensively through the rest paper.

In the sequel, we assume the parameter Θ is unknown and focus on the problem of approximating the sequence of probability measures

$$\mu_t(A) \triangleq \mathbb{P}_t\left\{\Theta \in A | Y_{1:t} = y_{1:t}\right\}, \quad t = 0, 1, 2, ..., \text{ where } A \in \mathcal{B}(D_\theta)$$

that result from the state–space Markov model and the sequence of observations $\{y_{1:t}\}_{t\geq 1}$.

3 Nested particle filtering algorithm

3.1 Recursive decomposition of μ_t

Assume that the observations $Y_{1:t-1} = y_{1:t-1}$ are fixed and let

$$\upsilon_{t,\theta}(A) = \mathbb{P}_t\left\{Y_t \in A | Y_{1:t-1} = y_{1:t-1}, \Theta = \theta\right\}, \quad A \in \mathcal{B}(\mathbb{R}^{d_y}), \tag{6}$$

be the probability measure associated to the (random) observation Y_t given $Y_{1:t-1} = y_{1:t-1}$ and the parameter vector $\Theta = \theta$. Let us assume that $\upsilon_{t,\theta}$ has a density $u_{t,\theta} : \mathbb{R}^{d_y} \to [0, +\infty)$ w.r.t. the Lebesgue measure, i.e., for any $A \in \mathcal{B}(\mathbb{R}^{d_y})$,

$$\upsilon_{t,\theta}(A) = \int I_A(y) u_{t,\theta}(y) dy.$$

The posterior probability measure of the parameter, μ_t, can be related to the predictive measure $\xi_{t,\theta}$ by way of the pdf $u_{t,\theta}(y)$. To be precise, for given $Y_t = y_t$ and $\Theta = \theta$, the density $u_{t,\theta}(y_t)$ can be written as the integral

$$u_{t,\theta}(y_t) = (g_{t,\theta}^{y_t}, \xi_{t,\theta}),$$

which yields the marginal likelihood of the parameter value θ, denoted in the sequel as

$$u_t(\theta) \triangleq u_{t,\theta}(y_t) = (g_{t,\theta}^{y_t}, \xi_{t,\theta}).$$

Then, it is a straightforward application of Bayes' theorem to show that the sequence of measures μ_t obeys the recursion

$$(h, \mu_t) = \frac{(hu_t, \mu_{t-1})}{(u_t, \mu_{t-1})}, \quad \text{for } t = 1, 2, ... \tag{7}$$

for any integrable function $h : D_\theta \to \mathbb{R}$.

Equation (7) suggests the implementation of a sequential Monte Carlo (SMC) approximation of μ_t. In particular, at time t one could

- draw N i.i.d. samples $\{\bar{\theta}_t^{(i)}\}_{1\leq i\leq N}$ from the posterior measure at time $t-1$, μ_{t-1},
- and then compute normalised importance weights proportional to the marginal likelihoods $u_t(\bar{\theta}_t^{(i)})$.

However, neither sampling from μ_{t-1} nor the computation of the likelihood $u_t(\theta)$ can be carried out exactly, hence some approximations are needed. This is explored in Subsections 3.2 and 3.3, respectively.

3.2 Sampling in the parameter space

Assume that a particle approximation $\mu_{t-1}^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_{t-1}^{(i)}}$ of μ_{t-1} is available. A natural way to generate a new sample of size N distributed approximately as μ_{t-1} is to *jitter* the particles $\theta_{t-1}^{(1)}, ..., \theta_{t-1}^{(N)}$.

Remark 2 *This random jittering, or rejuvenation, of the particles in the parameter space is necessary in order to avoid the degeneracy of the SMC method [24], but the error introduced by this step should be controlled. In the SMC^2 framework of [6], this is done by applying a particle Markov chain Monte Carlo (pMCMC) kernel to the particle set $\{\theta_{t-1}^{(i)}\}_{i=1}^N$ that leaves its underlying distribution invariant. However, this procedure implies the processing of the complete sequence of observations up to time t, $\mathbf{y}_{1:t}$, and, therefore, prevents a recursive implementation.*

To circumvent the drawback described in Remark 2, we propose to use Markov kernels of the form

$$\kappa_{N,\mathsf{p}}^{\theta_{t-1}^{(i)}}(d\theta) = (1-\epsilon_{N,\mathsf{p}})\delta_{\theta_{t-1}^{(i)}}(d\theta) + \epsilon_{N,\mathsf{p}}\bar{\kappa}^{\theta_{t-1}^{(i)}}(\theta)d\theta, \quad i = 1, 2, \ldots, N, \tag{8}$$

where $\epsilon_{N,\mathsf{p}} \in \left(0, \frac{1}{N^{\frac{\mathsf{p}}{2}}}\right]$, $\mathsf{p} \geq 1$, and $\bar{\kappa}^{\theta'}(\theta)$ is a pdf w.r.t. the Lebesgue measure, independent of N, centred at θ' and with support in D_θ, i.e., $\int \theta\bar{\kappa}^{\theta'}(\theta)d\theta = \theta'$ and $\int I_{D_\theta}(\theta)\bar{\kappa}^{\theta'}(\theta)d\theta = 1$. It is relatively straightforward to show that kernels in the class described by (8) satisfy the inequalities stated below.

Proposition 1 *If $\kappa_{N,\mathsf{p}}$ is selected as in Eq. (8), then*

$$\sup_{\theta' \in D_\theta} \int |h(\theta) - h(\theta')|\, \kappa_{N,\mathsf{p}}^{\theta'}(d\theta) \leq \frac{2\|h\|_\infty}{\sqrt{N}} \tag{9}$$

for any $h \in B(D_\theta)$, and

$$\sup_{\theta' \in D_\theta} \int \|\theta - \theta'\|^{\mathsf{p}}\, \kappa_{N,\mathsf{p}}^{\theta'}(d\theta) \leq \frac{c_\kappa^{\mathsf{p}}}{N^{\frac{\mathsf{p}}{2}}}, \tag{10}$$

where $c_\kappa < \infty$ is a constant independent of N.

Proof: It is straightforward. Simply note that $|h(\theta) - h(\theta')| \leq 2\|h\|_\infty$ to arrive at (9). Inequality (10) is readily obtained, with $c_\kappa = \sup_{\theta,\theta' \in D_\theta} \|\theta - \theta'\| < \infty$, if we recall that D_θ is defined to be compact. □

See [10, Section 5.1] for a more detailed discussion of the choice of the jittering kernel, including some variations on the family of equation (8). In the sequel, we assume that $\kappa_{N,\mathsf{p}}^{\theta_{t-1}^{(i)}}(d\theta)$ is selected according to (8), so that Proposition 1 holds.

3.3 Approximation of the parameter likelihood function $u_t(\theta)$

The second ingredient that we need in order to construct a SMC algorithm that approximates the measures μ_t is a method to compute the likelihood $u_t(\theta)$. For fixed $\Theta = \bar{\theta}_t^{(i)}$, the value $u_t(\bar{\theta}_t^{(i)})$ can be estimated using a standard particle filter (or *bootstrap* filter [16], see also [13]). This classical algorithm can be written down (in a convenient form) using the following notation for two random transformations of discrete sample sets on the state space $\mathcal{X}$.

Definition 1 *Let $\{x^{(j)}\}_{1\leq j\leq M}$ be a set of M points on $\mathcal{X}$. The random set*

$$\{\bar{x}^{(j)}\}_{1\leq j\leq M} = \Upsilon_{n,\theta}\left(\{x^{(j)}\}_{1\leq j\leq M}\right)$$

is obtained by sampling each $\bar{x}^{(j)}$ from the corresponding transition kernel $\tau_{n,\theta}(dx|x^{(j)})$, for $j = 1, ..., M$.

Definition 2 *Let $\{\bar{x}^{(j)}\}_{1\leq j\leq M}$ be a set of M points in $\mathcal{X}$. The set*

$$\{x^{(j)}\}_{1\leq j\leq M} = \Upsilon^{y_n}_{n,\theta}\left(\{\bar{x}^{(j)}\}_{1\leq j\leq M}\right)$$

is obtained by

- *computing normalised weights proportional to the likelihoods,*

$$v_n^{(j)} = \frac{g^{y_n}_{n,\theta}(\bar{x}_n^{(j)})}{\sum_{k=1}^{M} g^{y_n}_{n,\theta}(\bar{x}_n^{(k)})}, \quad j = 1, ..., M.$$

- *and then resampling with replacement the set $\{\bar{x}^{(j)}\}_{1\leq j\leq M}$ according to the weights $\{v_n^{(j)}\}_{1\leq j\leq M}$, i.e., assigning $x^{(j)} = \bar{x}^{(k)}$ with probability $v^{(k)}$, for $j = 1, ..., M$ and $k \in \{1, ..., M\}$.*

The standard particle filter, with M particles per time step and conditional on $\Theta = \theta_t^{(i)}$, can be outlined as follows.

Algorithm 1 *Bootstrap filter conditional on $\Theta = \theta_t^{(i)}$.*

1. Initialisation. *Draw M i.i.d. samples $x_0^{(i,j)}$, $j = 1, ..., M$, from the prior distribution τ_0.*
2. Recursive step. *Let $\{x_{n-1}^{(i,j)}\}_{1\leq j\leq M}$ be the set of available samples at time $n-1$, with $n \leq t$. The particle set is updated at time n in two steps:*
   (a) Compute $\{\bar{x}_n^{(i,j)}\}_{1\leq j\leq M} = \Upsilon_{n,\theta_t^{(i)}}\left(\{x_{n-1}^{(i,j)}\}_{1\leq j\leq M}\right)$.
   (b) Compute $\{x_n^{(i,j)}\}_{1\leq j\leq M} = \Upsilon^{y_n}_{n,\theta_t^{(i)}}\left(\{\bar{x}_n^{(i,j)}\}_{1\leq j\leq M}\right)$.

For $n = t$, we obtain random discrete approximations of the posterior probability measures $\xi_{t,\bar{\theta}_t^{(i)}}(dx_t)$ and $\phi_{t,\bar{\theta}_t^{(i)}}(dx_t)$ of the form

$$\xi^M_{t,\bar{\theta}_t^{(i)}}(dx_t) = \frac{1}{M}\sum_{j=1}^{M}\delta_{\bar{x}_t^{(i,j)}}(dx_t) \quad \text{and} \quad \phi^M_{t,\bar{\theta}_t^{(i)}}(dx_t) = \frac{1}{M}\sum_{j=1}^{M}\delta_{x_t^{(i,j)}}(dx_t), \tag{11}$$

respectively. Hence, the parameter likelihood $u_t(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi_{t,\bar{\theta}_t^{(i)}})$, which in general does not have a closed form solution, admits the Monte Carlo approximation

$$u_t^M(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi^M_{t,\bar{\theta}_t^{(i)}}) = \frac{1}{M}\sum_{j=1}^{M} g^{y_t}_{t,\bar{\theta}_t^{(i)}}(\bar{x}_t^{(i,j)}). \tag{12}$$

3.4 Nested particle filtering algorithm

We are now ready to describe the nested particle filtering algorithm which is the main object of analysis in this paper. Essentially, it is a recursive Monte Carlo filter on the parameter space D_θ that uses conditional bootstrap filters on $\mathcal{X}$ to approximate the parameter likelihoods. The algorithm is described below.

Algorithm 2 *Recursive algorithm for the particle approximation of μ_t, $t = 0, 1, 2, ...$*

1. Initialisation. *Draw N i.i.d. samples $\{\theta_0^{(i)}\}_{1\leq i\leq N}$ from the prior distribution $\mu_0(d\theta)$ and NM i.i.d. samples $\{x_0^{(i,j)}\}_{1\leq i\leq N; 1\leq j\leq M}$ from the prior distribution τ_0.*
2. Recursive step. *For $t \geq 1$, assume the particle set $\left\{\theta_{t-1}^{(i)}, \{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}\right\}_{1\leq i\leq N}$ is available and update it taking the following steps.*
   (a) *For each $i = 1, ..., N$*
      - *draw $\bar{\theta}_t^{(i)}$ from $\kappa_{N,\mathsf{p}}^{\theta_{t-1}^{(i)}}(d\theta)$,*
      - *update $\{\bar{x}_t^{(i,j)}\}_{1\leq j\leq M} = \Upsilon_{t,\bar{\theta}_t^{(i)}}\left(\{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}\right)$ and construct $\xi^M_{t,\bar{\theta}_t^{(i)}} = \frac{1}{M}\sum_{j=1}^M \delta_{\bar{x}_t^{(i,j)}}$,*
      - *compute the approximate likelihood $u_t^M(\bar{\theta}_t^{(i)}) = (g^{y_t}_{t,\bar{\theta}_t^{(i)}}, \xi^M_{t,\bar{\theta}_t^{(i)}})$, and*
      - *update the particle set $\{\tilde{x}_t^{(i,j)}\}_{1\leq j\leq M} = \Upsilon^{y_t}_{t,\bar{\theta}_t^{(i)}}\left(\{\bar{x}_t^{(i,j)}\}_{1\leq j\leq M}\right)$.*
   (b) *Compute normalised weights $w_t^{(i)} \propto u_t^M(\bar{\theta}_t^{(i)})$, $i = 1, ..., N$.*
   (c) *Resample: for each $i = 1, ..., N$, set $\left\{\theta_t^{(i)}, x_t^{(i,j)}\right\}_{1\leq j\leq M} = \left\{\bar{\theta}_t^{(l)}, \tilde{x}_t^{(l,j)}\right\}_{1\leq j\leq M}$ with probability $w_t^{(l)}$, where $l \in \{1, ..., N\}$.*

Step 2(a) in Algorithm 2 involves jittering the samples in the parameter space and then taking a single recursive step of a bank of N standard particle filters. In particular, for each $\bar{\theta}_t^{(i)}$, $1 \leq i \leq N$, we have to propagate and resample the particles $\{x_{t-1}^{(i,j)}\}_{1\leq j\leq M}$ so as to obtain a new set $\{\tilde{x}_t^{(i,j)}\}_{1\leq j\leq M}$.

Remark 3 *The cost of the recursive step in Algorithm 2 is independent of t. We only have to carry out regular 'prediction' and 'update' operations in a bank of standard particle filters. Hence, Algorithm 2 is sequential, purely recursive and can be implemented online. This is in contrast with the non-recursive (but otherwise similar) SMC^2 method of [6]. A detailed comparison of both techniques is presented in [10].*

Remark 4 *Algorithm 2 yields several Monte Carlo approximations. After the jittering step, we obtain the measure*

$$\bar{\mu}_{t-1}^{N,M} = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{\theta}_t^{(i)}}$$

which is an approximation of μ_{t-1} computed at time t. After the weights are computed at step 2(b), we have the neasure

$$\tilde{\mu}_t^{N,M} = \sum_{i=1}^N w_t^{(i)} \delta_{\bar{\theta}_t^{(i)}},$$

which approximates the posterior μ_t. After the resampling step 2(c) we have the (unweighted) approximation

$$\mu_t^{N,M} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_t^{(i)}}$$

of μ_t. Conditional predictive and filter measures on the state space are also computed by the inner filters, namely

$$\xi_{t,\bar{\theta}_t^{(i)}}^{M} = \frac{1}{M}\sum_{j=1}^{M} \delta_{\bar{x}_t^{(i,j)}} \quad \textit{and} \quad \phi_{t,\theta_t^{(i)}}^{M} = \frac{1}{M}\sum_{j=1}^{M} \delta_{x_t^{(i,j)}}.$$

4 Summary of theoretical results

In the rest of this paper we look into the particle approximations of the sequence produced by Algorithm 2. For notational simplicity, we assume that the numbers of particles in the inner and outer filters coincide, i.e., $N = M$. Thus, the approximation of the predictive measure $\xi_{t,\bar{\theta}_t^{(i)}}$ and the filter measure $\phi_{t,\theta_t^{(i)}}$ become $\xi_{t,\bar{\theta}_t^{(i)}}^{N}$ and $\phi_{t,\theta_t^{(i)}}^{N} = \frac{1}{N}\sum_{j=1}^{N} \delta_{x_t^{(i,j)}}$, respectively. For conciseness, we will also write

$$\bar{\mu}_t^N = \bar{\mu}_t^{N,N}, \quad \tilde{\mu}_t^N = \tilde{\mu}_t^{N,N} \quad \text{and} \quad \mu_t^N = \mu_t^{N,N}.$$

The complexity of Algorithm 2 with $N = M$ and a sequence of observations of length T, $Y_{1:T} = y_{1:T}$, becomes $\mathcal{O}(N^2 T)$ [10].

While in [10] we address the consistency of Algorithm 2 (as $N, M \to \infty$) for a finite-length sequence of observations, here we tackle the problem of proving that the proposed nested particle filter actually converges *uniformly over time* when the state space model satisfies a set of sufficient conditions. In particular, for the analysis in this paper we assume that

(i) the sequence of probability measures $\{\mu_t\}_{t\geq 0}$ is stable w.r.t. its initial value,

(ii) the Markov kernels $\tau_{t,\theta}(dx_t|x_{t-1})$ are mixing (uniformly, for all $\theta \in D_\theta$) and the likelihood functions $g_{t,\theta}^{y_t}(x_t)$ are normalised and bounded away from 0,

(iii) every Markov kernel $\tau_{t,\theta}(dx_t|x_{t-1})$ has an associated pdf w.r.t. the Lebesgue measure, denoted $\tau_{t,\theta}^{x_{t-1}}(x_t)$, and both these pdf's and the likelihood functions $g_{t,\theta}^{y_t}(x_t)$ are Lipschitz continuous w.r.t. the parameter θ.

These assumptions are made explicit in Section 5.1; then, in Sections 5.2 and 5.3 we progress toward the main result in this paper, which can be outlined as follows.

Result 1 *(Theorem 1, Section 5.3). If the assumptions (i), (ii) and (iii) above hold, and the jittering step of Algorithm 2 is implemented using the kernel $\kappa_{N,\mathsf{p}}$ defined in* (8)*, then*

$$\sup_{t\geq 0} \|(h, \mu_t^N) - (h, \mu_t)\|_p \leq r(N)$$

for every $h \in B(D_\theta)$ *and* $1 \leq p \leq \mathsf{p}$*, where* $r(N)$ *is a rate function (to be given explicitly) such that* $\lim_{N\to\infty} r(N) = 0$.

Result 1 has some relevant consequences. In particular, in Section 5.4 we prove that, under the same regularity assumptions on the state-space model, it is possible to "identify" the static parameter Θ, i.e., to compute estimates which are asymptotically exact.

Result 2 *(Theorem 2, Section 5.4). If* $\lim_{t\to\infty} \mu_t = \delta_{\theta_*}$ *for some* $\theta_* \in D_\theta$*, then*

$$\limsup_{t\to\infty} E\left[d(\mu_t^N, \delta_{\theta_*})\right] \leq \bar{r}(N),$$

where

- $d : \mathcal{P}(D_\theta) \times \mathcal{P}(D_\theta) \to [0, +\infty)$ *is a distance between probability measures, to be precisely defined in Section 5.4, and*
- $\bar{r}(N)$ *is a rate function (to be explicitly given) such that* $\lim_{N\to\infty} \bar{r}(N) = 0$.

5 Uniform convergence over time

In this section we carry out the analysis leading to the uniform convergence over time of the approximation errors $\|(h, \mu_t^N) - (h, \mu_t)\|_p$, the explicit derivation of error rates and the asymptotically exact estimation of Θ (under regularity assumptions on the sequence $\{\mu_t\}_{t\geq 0}$). Our argument is based on the approaches in [12] and [21], which rely on the stability of the sequences of measures to be approximated and the contractivity (under regularity assumptions) of the Markov kernels $\tau_{t,\tau}$.

Within this setup, we show the uniform convergence of the particle filters in the inner layer (i.e., conditional on the value of the parameter) and then establish the same result for the complete Algorithm 2. This leads naturally to Result 2 on the asymptotically exact estimation of the static parameters.

5.1 Notation and assumptions

5.1.1 Maps on the space of probability measures

Recall that $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(D_\theta)$ denote the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $(D_\theta, \mathcal{B}(D_\theta))$, respectively. We introduce the map $\Psi_t^\theta : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ that takes the predictive measure at time t into the predictive measure at time $t+1$. A precise definition si given below.

Definition 3 *For any integrable function* $f : \mathcal{X} \to \mathbb{R}$*, any time* $t \geq 0$ *and any parameter vector* $\theta \in D_\theta$*, we define the map* $\Psi_t^\theta : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ *as*

$$\left(f, \Psi_t^\theta(\alpha)\right) \triangleq \frac{\left(g_{t-1,\theta}^{y_{t-1}}(f, \tau_{t,\theta}), \alpha\right)}{\left(g_{t-1,\theta}^{y_{t-1}}, \alpha\right)}. \tag{13}$$

It is simple to check (e.g., by way of Eqs. (4) and (5)) that $\xi_{t,\theta} = \Psi_t^\theta(\xi_{t-1,\theta})$ for $t \geq 2$. In order to define Ψ_1^θ in a consistent manner, let us introduce

$$\begin{aligned} \phi_{-1} &\equiv \text{the uniform measure on } \mathcal{X}, \text{ and} \\ g_{0,\theta}^{y_0}(x) &= g_0(x) \triangleq 1 \quad \forall x \in \mathcal{X}. \end{aligned}$$

Then, $\xi_0 = \phi_0 = \tau_0$ (independently of θ) and $\xi_{1,\theta} = \Psi_1^\theta(\xi_0)$. Moreover, for any $0 \leq k \leq t$, let

$$\Psi_{t|k}^\theta \triangleq \Psi_t^\theta \circ \Psi_{t-1}^\theta \circ \cdots \circ \Psi_{k+1}^\theta,$$

where $\circ$ denotes composition. Note that $\Psi_{t|t-1}^\theta = \Psi_t^\theta$ and we adopt the convention $\Psi_{t|t}^\theta(\alpha) = \alpha$.

Definition 4 *For any integrable function $h : D_\theta \to \mathbb{R}$, any time $t > 0$ and any $\alpha \in \mathcal{P}(D_\theta)$, we define the map $\Lambda_t : \mathcal{P}(D_\theta) \to \mathcal{P}(D_\theta)$ as*

$$(h, \Lambda_t(\alpha)) \triangleq \frac{(hu_t, \alpha)}{(u_t, \alpha)},$$

hence $\mu_t = \Lambda_t(\mu_{t-1})$.

The composition $\Lambda_{t|k} = \Lambda_t \circ \cdots \circ \Lambda_{k+1}$ is constructed in the same way as for $\Psi_{t|k}^\theta$.

5.1.2 Stability of the posterior probability measures

Uniform convergence of particle filters over time can be guaranteed when the corresponding optimal filters satisfy some stability conditions [12]. In a similar manner, here we adopt stability assumptions for the sequence of posterior probability measures (in $\mathcal{P}(D_\theta)$) generated by the maps Λ_t, $t \geq 0$. These are made explicit below.

A. 1 *Let $\{y_t\}_{t\geq 1}$ be an arbitrary sequence of observations and let*

$$\mathcal{S}(h,T) = \sup_{\alpha,\eta\in\mathcal{P}(D_\theta);k\geq 0} \left|\left(h, \Lambda_{k+T|k}(\alpha)\right) - \left(h, \Lambda_{k+T|k}(\eta)\right)\right|,$$

where $h : D_\theta \to \mathbb{R}$. Then, $\lim_{T\to\infty} \mathcal{S}(h,T) = 0$ for every $h \in B(D_\theta)$.

A. 2 *For every $h \in B(D_\theta)$ there exist real constants $\bar{b}_1 > 0$ and $\bar{b}_2 > 0$, and a natural constant $\bar{T}_0 \geq 1$, such that*

$$\mathcal{S}(h,T) \leq \bar{b}_1 \exp\left\{-\bar{b}_2 T\right\} \quad \textit{for every } T \geq \bar{T}_0.$$

5.1.3 Bounds and Lipschitz continuity

The latter stability assumptions for the maps Λ_t are combined with the existence of certain bounds for the family of likelihood functions $g_{t,\theta}^{y_t}$ and Markov kernels $\tau_{t,\theta}$. These assumptions are made to ensure that the optimal inner filters (conditional on θ) are stable for any choice of the parameters within the support D_θ and their particle approximations converge uniformly over time. They correspond to similar standard assumptions, e.g., in [11] or [21], used in the analysis of conventional particle filters.

A. 3 *Let $\{y_t\}_{t\geq 1}$ be an arbitrary but fixed sequence of observations. The likelihood functions are normalised and bounded away from 0, i.e., there exists a positive constant $a < \infty$ such that*

$$\inf_{x\in\mathcal{X},\theta\in D_\theta,t\geq 1} g_{t,\theta}^{y_t}(x) \geq \frac{1}{a} \quad \textit{and} \quad \sup_{x\in\mathcal{X},\theta\in D_\theta,t\geq 1} g_{t,\theta}^{y_t}(x) \leq 1.$$

Let $\tau_{t+m|t,\theta}(dx_{t+m}|x_t)$ denote the composition of m consecutive Markov kernels, from time $t+1$ to time $t+m$, with starting point $x_t \in \mathcal{X}$ at time t. In particular, the integral of a function $f : \mathcal{X} \to \mathbb{R}$ w.r.t. the composite kernel $\tau_{t+m|t,\theta}(dx_{t+m}|x_t)$ can be explicitly written as

$$(f, \tau_{t+m|t,\theta}(\cdot|x_t)) \triangleq \int \cdots \int f(x_{t+m})\tau_{t+m,\theta}(dx_{t+m}|x_{t+m-1})\tau_{t+m-1,\theta}(dx_{t+m-1}|x_{t+m-2}) \cdots \tau_{t+1,\theta}(dx_{t+1}|x_t).$$

We make the following assumption on the composition of kernels.

A. 4 *For a given integer $m > 0$ there exists a constant $0 < \epsilon_\tau < 1$ such that, for every Borel set $A \in \mathcal{B}(\mathcal{X})$,*

$$\inf_{t\geq 0,(x,x')\in\mathcal{X}^2,\theta\in D_\theta} \frac{\tau_{t+m|t,\theta}(A|x)}{\tau_{t+m|t,\theta}(A|x')} \geq \epsilon_\tau.$$

The jittering of the particles in the parameter space introduces a perturbation in the inner layer of particle filters of Algorithm 2. The procedure works when the effect of this perturbation on the approximating measures $\phi_{t,\theta}^N$ and $\xi_{t,\theta}^N$ is "sufficiently small", which can only be ensured when the corresponding measures enjoy some continuity property w.r.t. the parameters. This assumption is made explicit below.

A. 5 *Every Markov kernel $\tau_{t,\theta}(dx|x')$ has a density w.r.t. the Lebesgue measure, denoted $\tau_{t,\theta}^{x'}(dx)$. The functions $g_{t,\theta}^{y_t}(x)$ and $\tau_{t,\theta}^{x'}(x)$ are Lipschitz in the parameter θ for every $(x, x') \in \mathcal{X}^2$ and $t \geq 0$. In particular, there exists constants $L_g < \infty$ and L_τ such that, for any $\theta, \theta' \in D_\theta$,*

$$\begin{aligned}
\sup_{t\geq 1;x\in\mathcal{X}} |g_{t,\theta}^{y_t}(x) - g_{t,\theta'}^{y_t}(x)| &\leq L_g\|\theta - \theta'\|, \\
\sup_{t\geq 0;(x,x')\in\mathcal{X}^2} |\tau_{t,\theta}^{x'}(x) - \tau_{t,\theta'}^{x'}(x)| &\leq L_\tau\|\theta - \theta'\|.
\end{aligned}$$

Remark 5 *Let $L_{g,\tau} = L_g \vee L_\tau$. If assumptions A.5 and A.3 hold, then it is not difficult to show that*

$$|(f, \xi_{t,\theta}) - (f, \xi_{t,\theta'})| \leq ta^t\|f\|_\infty L_{g,\tau}\|\theta - \theta'\| \tag{14}$$

for any $f \in B(\mathcal{X})$ and $t \geq 1$, which corresponds to [10, Assumption A.3]. Integrals of the form $(f, \phi_{t,\theta})$ are also Lipschitz functions w.r.t. θ, since $(f, \xi_{t,\theta}) - (f, \xi_{t,\theta'}) = ((f, \tau_{t,\theta}), \phi_{t-1,\theta}) - ((f, \tau_{t,\theta'}), \phi_{t-1,\theta'})$.

5.1.4 An auxiliary result

For any pair of integers $0 < s < t$ we can explicitly construct the conditional pdf of the subsequence of observations $y_{s:t}$ given a point $X_s = x_s$ in the state space and a choice parameters $\Theta = \theta$. We denote

this density as $g^{y_{s:t}}_{s:t,\theta}(x_s)$, with the notation chosen to make explicit that, for fixed $y_{s:t}$, this is a function of the state value x_s (i.e., it is interpreted as a likelihood). It is not difficult to show that

$$g^{y_{s:t}}_{s:t,\theta}(x_s) = \int \cdots \int \prod_{j=s}^{t} g^{y_j}_{j,\theta}(x_j) \prod_{l=s+1}^{t} \tau_{l,\theta}(dx_l|x_{l-1}). \tag{15}$$

We also introduce a specific notation for the conditional distribution of the state X_j conditional on $X_{j-1} = x_{j-1}$, $\Theta = \theta$ and the subsequence of observations from time j up to time t, $y_{j:t}$. For any $j \leq t$, this is a Markov kernel, denoted $\mathsf{k}^{y_{j:t}}_{j,\theta}(dx_j|x_{j-1})$, that can be explicitly written as

$$\mathsf{k}^{y_{j:t}}_{j,\theta}(dx_j|x_{j-1}) = \frac{g^{y_{j:t}}_{j:t,\theta}(x_j)\tau_{j,\theta}(dx_j|x_{j-1})}{\int g^{y_{j:t}}_{j:t,\theta}(\tilde{x}_j)\tau_{j,\theta}(d\tilde{x}_j|x_{j-1})} \tag{16}$$

via the Bayes' theorem. If the observation sequence is fixed, then the composite probability measure

$$\mathsf{K}^{y_{s:t}}_{s:t+1,\theta}(dx_{t+1}|x_s) = \int \cdots \int \tau_{t+1,\theta}(dx_{t+1}|x_t) \prod_{j=s+1}^{t} \mathsf{k}^{y_{j:t}}_{j,\theta}(dx_j|x_{j-1}) \tag{17}$$

is a Markov kernel on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

The composite likelihood in (15) and the Markov kernel in (17) can be used to write integrals w.r.t. the composite map $\Psi^{\theta}_{t+1|s}$ explicitly. To be specific, given a probability measure $\alpha \in \mathcal{P}(\mathcal{X})$, it is an exercise to show that

$$\left(f, \Psi^{\theta}_{t+1|s}(\alpha)\right) = \frac{\left((f, \mathsf{K}^{y_{s:t}}_{s:t+1,\theta}) g^{y_{s:t}}_{s:t,\theta}, \alpha\right)}{\left(g^{y_{s:t}}_{s:t,\theta}, \alpha\right)}. \tag{18}$$

The representation in (18), together with assumptions A.3 and A.4, enables the application of standard results from [11] which become instrumental in the analysis of Algorithm 2.

We first define the Dobrushin contraction coefficient [12] for Markov kernels and then show how it can be used to control the difference between between two probability measures $\Psi^{\theta}_{t+1|s}(\alpha)$ and $\Psi^{\theta}_{t+1|s}(\eta)$ which are constructed using the same composite map $\Psi^{\theta}_{t+1|s}$ (and, in particular, the same observation subsequence $y_{s:t+1}$) but different initial conditions $\alpha \neq \eta$.

Definition 5 *The Dobrushin contraction coefficient of a Markov kernel K_θ from $\mathcal{X}$ onto $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is*

$$\beta(K_\theta) \triangleq \sup_{x,x' \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})} |K_\theta(A|x) - K_\theta(A|x')| \leq 1.$$

An upper bound for the contraction coefficient of the kernel $\mathsf{K}^{y_{s:t}}_{s:t+1,\theta}$, explicitly given in terms of the constants m, ϵ_τ and a in assumptions A.4 and A.3, is given below.

Lemma 1 *If assumptions A.3 and A.4 hold, then*

$$\beta(\mathsf{K}^{y_{s:t}}_{s:t+1,\theta}) \leq \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{t-s+1}{m} \rfloor} \tag{19}$$

for every $\theta \in D_\theta$.

Proof: Since the inequalities in A.3 and A.4 are assumed to hold uniformly over the parameter space D_θ, the bound in (19) follows readily from Proposition 4.3.3 in [11] (see also [11, Corollary 4.3.3]). □

From Lemma 1, and given a test function $f \in B(\mathcal{X})$, we can obtain a bound for the difference $\left|(f, \Psi^\theta_{t+1|s}(\alpha)) - (f, \Psi^\theta_{t+1|s}(\eta))\right|$ that will ease considerably the convergence analysis for Algorithm 2.

Lemma 2 *Assume that A.3 and A.4 hold true. Then, for any time indices $0 \leq s \leq t$, any pair of probability measures $\alpha, \eta \in \mathcal{P}(\mathcal{X})$ and any test function $f \in B(\mathcal{X})$ there exists another bounded function $\tilde{f}_s \in B(\mathcal{X})$, with $\|f\|_\infty \leq 1$, such that*

$$\left|(f, \Psi^\theta_{t+1|s}(\alpha)) - (f, \Psi^\theta_{t+1|s}(\eta))\right| \leq 2\|f\|_\infty \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{t-s+1}{m} \rfloor} \frac{a^m}{\epsilon_\tau} \left|(\tilde{f}_s, \alpha) - (\tilde{f}_s, \eta)\right|. \tag{20}$$

Proof: From [11, Proposition 4.3.7] we obtain an upper bound for the difference of integrals that depends on the Dobrushin coefficient of the Markov kernel $\mathsf{K}^{y_{s:t}}_{s:t+1,\theta}$, namely

$$\left|(f, \Psi^\theta_{t+1|s}(\alpha)) - (f, \Psi^\theta_{t+1|s}(\eta))\right| \leq 2\|f\|_\infty \beta(\mathsf{K}^{y_{s:t}}_{s:t+1,\theta}) \left(\sup_{x_s \in \mathcal{X}} \frac{g^{y_{s:t}}_{s:t,\theta}(x_s)}{(g^{y_{s:t}}_{s:t,\theta}, \alpha)}\right) \left|(\tilde{f}_s, \alpha) - (\tilde{f}_s, \eta)\right|, \tag{21}$$

for some $\tilde{f}_s : \mathcal{X} \to \mathbb{R}$ with $\|\tilde{f}_s\| \leq 1$. Moreover, from the definition of the composite likelihood in (15) and the assumption $g^{y_j}_{j,\theta} \leq 1$ for every $j \geq 1$ and $\theta \in D_\theta$ (in A.3), it follows that

$$g^{y_{s:t}}_{s:t,\theta}(x_s) \leq (g^{y_{s+m:t}}_{s+m:t,\theta}, \tau_{s+m|s,\theta}(\cdot|x_s)) \tag{22}$$

whereas, from the bound $g^{y_j}_{j,\theta}(x) \geq \frac{1}{a}$, for all $j \geq 1$ and $\theta \in D_\theta$ (in A.3) and the assumption A.4, we obtain that

$$(g^{y_{s:t}}_{s:t,\theta}, \alpha) \geq \frac{\epsilon_\tau}{a^m} (g^{y_{s+m:t}}_{s+m:t}, \tau_{s+m|s,\theta}(\cdot|\tilde{x}_s)) \tag{23}$$

for any $\tilde{x}_s \in \mathcal{X}$. In particular, for $x_s = \tilde{x}_s$, the inequalities (22) and (23) taken together yield

$$\frac{g^{y_{s:t}}_{s:t,\theta}(x_s)}{(g^{y_{s:t}}_{s:t,\theta}, \alpha)} \leq \frac{a^m}{\epsilon_\tau}$$

independently of x_s. This, in turn, enables us to rewrite (21) as

$$\left|(f, \Psi^\theta_{t+1|s}(\alpha)) - (f, \Psi^\theta_{t+1|s}(\eta))\right| \leq 2\|f\|_\infty \beta(\mathsf{K}^{y_{s:t}}_{s:t+1,\theta}) \frac{a^m}{\epsilon_\tau} \left|(\tilde{f}_s, \alpha) - (\tilde{f}_s, \eta)\right|. \tag{24}$$

By combining Lemma 1 with (24) we readily obtain the inequality (20) and complete the proof. □

5.2 Uniform convergence of the inner particle filters

We first establish the uniform convergence over time of a conditional bootstrap filter when the parameter corresponds to a Markov chain with the kernel $\kappa^{\theta'}_{N,\mathsf{p}}(d\theta)$ described in Section 3.2. To be specific, assume that the model is the same as in Section 2.2 (in particular, the parameter Θ is random but fixed) however we run a modification of Algorithm 1 where, at each time t, we generate a random variate θ_t with conditional probability measure $\kappa^{\theta_{t-1}}_{N,\mathsf{p}}(d\theta_t)$. The Markov chain is initialized with θ_0 drawn from the prior μ_0. The particle filter conditional on the chain $\{\theta_t\}_{t\geq 0}$ constructed in this manner is outlined below.

Algorithm 3 *Bootstrap filter conditional on a Markov chain of parameter realisations given by* $\theta_0 \sim \mu_0(d\theta)$ *and* $\theta_t \sim \kappa_{N,\mathsf{p}}^{\theta_{t-1}}(d\theta)$, $t \geq 1$.

1. Initialisation. *Draw* N *i.i.d. samples from* τ_0, *denoted* $x_0^{(i)}$, $i = 1, \ldots, N$.
2. Recursive step. *Let* $\{x_{t-1}^{(i)}\}_{1 \leq i \leq N}$ *be the particles generated at time* $t-1$. *At time* t, *proceed with the two steps below.*
   (a) *For* $i = 1, ..., N$, *draw a sample* $\bar{x}_t^{(i)}$ *from the probability distribution* $\tau_{t,\theta_t}(\cdot|x_{t-1}^{(i)})$ *and compute the normalised weight*
   $$w_t^{(i)} = \frac{g_{t,\theta_t}^{y_t}(\bar{x}_t^{(i)})}{\sum_{k=1}^{N} g_{t,\theta_t}^{y_t}(\bar{x}_t^{(k)})}. \tag{25}$$
   (b) *For* $i = 1, ..., N$, *let* $x_t^{(i)} = \bar{x}_t^{(k)}$ *with probability* $w_t^{(k)}$, $k \in \{1, ..., N\}$.

Note that, for any particle $\bar{\theta}_t^{(i)}$, $i \in \{1, ..., N\}$, at time t in the nested particle filter described by Algorithm 2, each conditional particle filter in the inner layer can be described as an instance of Algorithm 3. Indeed, by tracking the "history" of $\bar{\theta}_t^{(i)}$ across the resampling steps of Algorithm 2, we find that there is a sequence on D_θ of the form $\theta_{0|t}^{(i)}, \theta_{1|t}^{(i)}, \ldots, \theta_{t|t}^{(i)}$ such that,

- for $n = 0$, $\theta_{0|t}^{(i)}$ is drawn from μ_0,
- for any $0 \leq n \leq t$, $\theta_{n|t}^{(i)}$ is drawn from the kernel $\kappa_{N,\mathsf{p}}^{\theta_{n-1|t}^{(i)}}$ and,
- for $n = t$, $\theta_{t|t}^{(i)} = \bar{\theta}_t^{(i)}$.

Lemma 3 below states that the approximation (f, ξ_{t,θ_t}^N), where $f \in B(\mathcal{X})$, generated by Algorithm 3 actually converges to (f, ξ_{t,θ_t}), as N increases, uniformly over time under a subset of the assumptions in Section 5.1. This is a non-trivial result. Note that ξ_{t,θ_t} is the predictive probability measure at time t associated to the state space model $\{\tau_0, \tau_{n,\Theta}, g_{n,\Theta}^{y_n}\}_{1 \leq n \leq t}$, where $\Theta = \theta_t$ is fixed, while ξ_{t,θ_t}^N results from Algorithm 3, where the parameter value is effectively changing over time as a realisation $\theta_0, \theta_1, ..., \theta_t$ of a Markov chain up to time t.

Lemma 3 *Let* $\{\theta_t\}_{t \geq 0}$ *denote a Markov chain on the compact set* D_θ, *generated from the prior* μ_0 *and the kernels* $\kappa_{N,\mathsf{p}}^{\theta_{t-1}}(d\theta)$ *constructed as in Eq.* (8). *Let* $\xi_{t,\theta_t}^N = \frac{1}{N}\sum_{n=1}^{N} \delta_{\bar{x}_t^{(n)}}$ *be the sequence of approximate predictive measures generated by Algorithm 3. If assumptions A.3, A.4 and A.5 hold then there exists a real constant* $\bar{C}$, *independent of* N *and independent of the sequence* $\{\theta_t\}_{t \geq 0}$, *such that, for any* $f \in B(\mathcal{X})$ *and any* $1 \leq p \leq \mathsf{p}$,
$$\sup_{t \geq 0} \left\| (f, \xi_{t,\theta_t}^N) - (f, \xi_{t,\theta_t}) \right\|_p \leq \frac{\bar{C}}{\sqrt{N}}. \tag{26}$$
In particular, $\lim_{N \to \infty} \sup_{t \geq 0} \|(f, \xi_{t,\theta_t}^N) - (f, \xi_{t,\theta_t})\|_p = 0$.

Proof: We look into the approximation error $\left|(f,\xi^N_{t,\theta_t})-(f,\xi_{t,\theta_t})\right|$, which can be written as

$$\begin{aligned}\left|(f,\xi^N_{t,\theta_t})-(f,\xi_{t,\theta_t})\right| &= \left|\sum_{k=0}^{t-1}\left(f,\Psi^{\theta_t}_{t|t-k}\left(\xi^N_{t-k,\theta_{t-k}}\right)\right)-\left(f,\Psi^{\theta_t}_{t|t-k-1}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right.\\ &\quad\left.+\left(f,\Psi^{\theta_t}_{t|0}\left(\xi^N_{0,\theta_0}\right)\right)-\left(f,\Psi^{\theta_t}_{t|0}\left(\tau_0\right)\right)\right|\\ &\leq \sum_{k=0}^{t-1}\left|\left(f,\Psi^{\theta_t}_{t|t-k}\left(\xi^N_{t-k,\theta_{t-k}}\right)\right)-\left(f,\Psi^{\theta_t}_{t|t-k-1}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right|\\ &\quad+\left|\left(f,\Psi^{\theta_t}_{t|0}\left(\xi^N_{0,\theta_0}\right)\right)-\left(f,\Psi^{\theta_t}_{t|0}\left(\tau_0\right)\right)\right|, \qquad (27)\end{aligned}$$

where the equality follows from a 'telescopic' decomposition of the difference $(f,\xi^N_{t,\theta_t})-(f,\xi_{t,\theta_t})$. To see this, simply recall that $\xi^N_{0,\theta_0}\equiv\phi^N_{0,\theta_0}\equiv\tau^N_0$ (independently of θ_0 according to the model in Section 2.2) and note that $\Psi^{\theta_t}_{t|0}(\tau_0)=\xi_{t,\theta_t}$. By way of Minkowski's inequality, (27) enables us to express the L_p norm of the approximation error (for $p\geq 1$) as

$$\begin{aligned}\left\|(f,\xi^N_{t,\theta_t})-(f,\xi_{t,\theta_t})\right\|_p &\leq \sum_{k=0}^{t-1}\left\|\left(f,\Psi^{\theta_t}_{t|t-k}\left(\xi^N_{t-k,\theta_{t-k}}\right)\right)-\left(f,\Psi^{\theta_t}_{t|t-k-1}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right\|_p\\ &\quad+\left\|\left(f,\Psi^{\theta_t}_{t|0}\left(\xi^N_{0,\theta_0}\right)\right)-\left(f,\Psi^{\theta_t}_{t|0}\left(\tau_0\right)\right)\right\|_p, \qquad (28)\end{aligned}$$

The last term in the decomposition above can be easily upper bounded using Lemma 2, namely

$$\begin{aligned}\left\|\left(f,\Psi^{\theta_t}_{t|0}\left(\xi^N_{0,\theta_0}\right)\right)-\left(f,\Psi^{\theta_t}_{t|0}\left(\tau_0\right)\right)\right\|_p &\leq 2\|f\|_\infty\left(1-\frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor\frac{t}{m}\rfloor}\frac{a^m}{\epsilon_\tau}\left\|(\tilde{f}_0,\tau^N_0)-(\tilde{f}_0,\tau_0)\right\|_p,\\ &\leq 2\|f\|_\infty\left(1-\frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor\frac{t}{m}\rfloor}\frac{a^m}{\epsilon_\tau}\frac{\tilde{C}_0}{\sqrt{N}} \qquad (29)\end{aligned}$$

where $\|\tilde{f}_0\|_\infty\leq 1$ and the second inequality follows readily from the fact that $\tau^N_0=\xi^N_{0,\theta_0}$ is an i.i.d. Monte Carlo approximation of τ_0 (hence, $\tilde{C}_0<\infty$ is a constant independent of N). For the remaining terms in the sum of (28), Lemma 2 yields

$$\begin{aligned}&\left\|\left(f,\Psi^{\theta_t}_{t|t-k}\left(\xi^N_{t-k,\theta_{t-k}}\right)\right)-\left(f,\Psi^{\theta_t}_{t|t-k-1}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right\|_p \leq\\ &2\|f\|_\infty\left(1-\frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor\frac{k}{m}\rfloor}\frac{a^m}{\epsilon_\tau}\left\|\left(\tilde{f}_{t-k},\xi^N_{t-k,\theta_{t-k}}\right)-\left(\tilde{f}_{t-k},\Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right\|_p. \qquad (30)\end{aligned}$$

where $\|\tilde{f}_{t-k}\|_\infty\leq 1$.

In order to convert (30) into an explicit error rate, we need to derive bounds for errors of the form $\left\|\left(h,\xi^N_{t-k,\theta_{t-k}}\right)-\left(h,\Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right\|_p$, where $h:\mathcal{X}\to\mathbb{R}$ with $\|h\|_\infty\leq 1$. With this aim, we consider the triangular inequality

$$\begin{aligned}\left\|\left(h,\xi^N_{t-k,\theta_{t-k}}\right)-\left(h,\Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right\|_p &\leq \left\|\left(h,\xi^N_{t-k,\theta_{t-k}}\right)-E\left[\left(h,\xi^N_{t-k,\theta_{t-k}}\right)|\mathcal{G}_{t-k}\right]\right\|_p+\\ &\quad\left\|E\left[\left(h,\xi^N_{t-k,\theta_{t-k}}\right)|\mathcal{G}_{t-k}\right]-\left(h,\Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right\|_p, \qquad (31)\end{aligned}$$

where $\mathcal{G}_{t-k} = \sigma\left(x_{0:t-k-1}^{(n)}, \bar{x}_{1:t-k-1}^{(n)}, \{\theta_s\}_{s\geq 0}; 1 \leq n \leq N\right)$ is the σ-algebra generated by the random variables between brackets, and analyse the two terms on the right hand side separately.

For the first term on the right hand side of (31), we note that

$$\left(h, \xi_{t-k,\theta_{t-k}}^N\right) - E\left[\left(h, \xi_{t-k,\theta_{t-k}}^N\right)|\mathcal{G}_{t-k}\right] = \frac{1}{N}\sum_{n=1}^N \bar{S}_{t-k}^{(n)},$$

where

$$\bar{S}_{t-k}^{(n)} = h(\bar{x}_{t-k}^{(n)}) - E\left[h(\bar{x}_{t-k}^{(n)})|\mathcal{G}_{t-k}\right], \quad n = 1, ..., N,$$

are zero-mean and conditionally (on $\mathcal{G}_{t-k}$) independent r.v.'s. Therefore it is straightforward to show that

$$E\left[\left|\left(h, \xi_{t-k,\theta_{t-k}}^N\right) - E\left[\left(h, \xi_{t-k,\theta_{t-k}}^N\right)|\mathcal{G}_{t-k}\right]\right|^p |\mathcal{G}_{t-k}\right] = E\left[\left|\frac{1}{N}\sum_{n=1}^N \bar{S}_{t-k}^{(n)}\right|^p |\mathcal{G}_{t-k}\right] \leq \frac{c^p}{N^{\frac{p}{2}}} \tag{32}$$

for some constant $c > 0$ independent of N and independent of the distribution of the variables $\bar{S}_{t-k}^{(n)}$, $n = 1, ..., N$ (in particular, independent of the sequence $\{\theta_t\}_{t\geq 0}$). Taking expectations on both sides of (32), and then exponentiating by $\frac{1}{p}$, yields

$$\left\|\left(h, \xi_{t-k,\theta_{t-k}}^N\right) - E\left[\left(h, \xi_{t-k,\theta_{t-k}}^N\right)|\mathcal{G}_{t-k}\right]\right\|_p \leq \frac{c}{\sqrt{N}}. \tag{33}$$

To find a rate for the second term in (31), we note that

$$E\left[\left(h, \xi_{t-k,\theta_{t-k}}^N\right)|\mathcal{G}_{t-k}\right] = \frac{\left(g_{t-k-1,\theta_{t-k-1}}^{y_{t-k-1}}\left(h, \tau_{t-k,\theta_{t-k}}\right), \xi_{t-k-1,\theta_{t-k-1}}^N\right)}{\left(g_{t-k-1,\theta_{t-k-1}}^{y_{t-k-1}}, \xi_{t-k-1,\theta_{t-k-1}}^N\right)} \tag{34}$$

whereas

$$\left(h, \Psi_{t-k}^{\theta_t}\left(\xi_{t-k-1,\theta_{t-k-1}}^N\right)\right) = \frac{\left(g_{t-k-1,\theta_t}^{y_{t-k-1}}\left(h, \tau_{t-k,\theta_t}\right), \xi_{t-k-1,\theta_{t-k-1}}^N\right)}{\left(g_{t-k-1,\theta_t}^{y_{t-k-1}}, \xi_{t-k-1,\theta_{t-k-1}}^N\right)}. \tag{35}$$

Subtracting (35) from (34) and then rearranging terms yields

$$\begin{aligned} E\left[\left(h, \xi_{t-k,\theta_{t-k}}^N\right)|\mathcal{G}_{t-k}\right] - \left(h, \Psi_{t-k}^{\theta_t}\left(\xi_{t-k-1,\theta_{t-k-1}}^N\right)\right) &= \\ \frac{\left(g_{t-k-1,\theta_{t-k-1}}^{y_{t-k-1}}\left(h, \tau_{t-k,\theta_{t-k}}\right) - g_{t-k-1,\theta_t}^{y_{t-k-1}}\left(h, \tau_{t-k,\theta_t}\right), \xi_{t-k-1,\theta_{t-k-1}}^N\right)}{\left(g_{t-k-1,\theta_t}^{y_{t-k-1}}, \xi_{t-k-1,\theta_{t-k-1}}^N\right)} &+ \\ \frac{E\left[\left(h, \xi_{t-k,\theta_{t-k}}^N\right)|\mathcal{G}_{t-k}\right] \times \left(g_{t-k-1,\theta_t}^{y_{t-k-1}} - g_{t-k-1,\theta_{t-k-1}}^{y_{t-k-1}}, \xi_{t-k-1,\theta_{t-k-1}}^N\right)}{\left(g_{t-k-1,\theta_t}^{y_{t-k-1}}, \xi_{t-k-1,\theta_{t-k-1}}^N\right)}, & \end{aligned}$$

hence

$$\begin{aligned} \left|E\left[\left(h, \xi_{t-k,\theta_{t-k}}^N\right)|\mathcal{G}_{t-k}\right] - \left(h, \Psi_{t-k}^{\theta_t}\left(\xi_{t-k-1,\theta_{t-k-1}}^N\right)\right)\right| &\leq \\ a \times \left(\left|g_{t-k-1,\theta_{t-k-1}}^{y_{t-k-1}}\left(h, \tau_{t-k,\theta_{t-k}}\right) - g_{t-k-1,\theta_t}^{y_{t-k-1}}\left(h, \tau_{t-k,\theta_t}\right)\right|, \xi_{t-k-1,\theta_{t-k-1}}^N\right) &+ \\ a \times \left(\left|g_{t-k-1,\theta_t}^{y_{t-k-1}} - g_{t-k-1,\theta_{t-k-1}}^{y_{t-k-1}}\right|, \xi_{t-k-1,\theta_{t-k-1}}^N\right), & \end{aligned} \tag{36}$$

where we have used the obvious bounds $E\left[\left(h,\xi^N_{t-k,\theta_{t-k}}\right)|\mathcal{G}_{t-k}\right] \leq \|h\|_\infty \leq 1$ and, from assumption A.3, $\left(g^{y_{t-k-1}}_{t-k-1,\theta_t}, \xi^N_{t-k-1,\theta_{t-k-1}}\right) \geq a^{-1}$.

From assumption A.5, the likelihoods $g^{y_t}_{t,\theta}(x)$ are Lipschitz in the parameter θ, with constant L_g independent of t and x. In particular,

$$\sup_{x\in\mathcal{X},t\geq T}\left|g^{y_{t-k-1}}_{t-k-1,\theta_t}(x) - g^{y_{t-k-1}}_{t-k-1,\theta_{t-k-1}}(x)\right| \leq L_g\|\theta_t - \theta_{t-k-1}\|. \tag{37}$$

Also from assumption A.5, the kernels $\tau_{t,\theta}(dx|x) \in \mathcal{P}(\mathcal{X})$ are endowed with densities w.r.t. the Lebesgue measure, hence we can write

$$\begin{aligned}\left|g^{y_{t-k-1}}_{t-k-1,\theta_{t-k-1}}(x)\left(h,\tau_{t-k,\theta_{t-k}}\right)(x) - g^{y_{t-k-1}}_{t-k-1,\theta_t}(x)\left(h,\tau_{t-k,\theta_t}\right)(x)\right| &= \\ \left|g^{y_{t-k-1}}_{t-k-1,\theta_{t-k-1}}(x)\int h(x')\tau^x_{t-k,\theta_{t-k}}(x')dx' - g^{y_{t-k-1}}_{t-k-1,\theta_t}(x)\int h(x')\tau^x_{t-k,\theta_t}(x')dx'\right|&\end{aligned}$$

and a simple triangle inequality yields

$$\begin{aligned}\left|g^{y_{t-k-1}}_{t-k-1,\theta_{t-k-1}}(x)\left(h,\tau_{t-k,\theta_{t-k}}\right)(x) - g^{y_{t-k-1}}_{t-k-1,\theta_t}(x)\left(h,\tau_{t-k,\theta_t}\right)(x)\right| &\leq \\ \left|\left(g^{y_{t-k-1}}_{t-k-1,\theta_{t-k-1}}(x) - g^{y_{t-k-1}}_{t-k-1,\theta_{t-k}}(x)\right)\int h(x')\tau^x_{t-k,\theta_{t-k}}(x')dx'\right| + &\\ \left|\int h(x')\left(g^{y_{t-k-1}}_{t-k-1,\theta_{t-k}}(x)\tau^x_{t-k,\theta_{t-k}}(x') - g^{y_{t-k-1}}_{t-k-1,\theta_t}(x)\tau^x_{t-k,\theta_t}(x')\right)dx'\right| &\leq \\ L_g \vee L_{g,\tau}\left(\|\theta_{t-k-1} - \theta_{t-k}\| + \|\theta_t - \theta_{t-k}\|\right),&\end{aligned} \tag{38}$$

where the second inequality is satisfied because the product $g^{y_t}_{t,\theta}\tau^x_{t,\theta'}(x')$ is Lipschitz in θ for every $t \geq 1$ and $x, x' \in \mathcal{X}$ (a consequence of assumption A.5) with constant $L_{g,\tau}$.

If we substitute (37) and (38) back into (36) we obtain

$$\left|E\left[\left(h,\xi^N_{t-k,\theta_{t-k}}\right)|\mathcal{G}_{t-k}\right] - \left(h,\Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right| \leq 2aL\sum_{j=0}^{k}\|\theta_{t-j} - \theta_{t-j-1}\| \tag{39}$$

where we have introduced the constant $L = \max\{L_g, L_{g,\tau}\}$ and taken advantage of the straightforward inequality $\|\theta_t - \theta_{t-k-1}\| \leq \sum_{j=0}^{k}\|\theta_{t-j} - \theta_{t-j-1}\|$. Raising both sides of (39) to power p and then taking expectations yields

$$\begin{aligned}E\left[\left|E\left[\left(h,\xi^N_{t-k,\theta_{t-k}}\right)|\mathcal{G}_{t-k}\right] - \left(h,\Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right|^p\right] &\leq (2aL)^p E\left[\left|\sum_{j=0}^{k}\|\theta_{t-j} - \theta_{t-j-1}\|\right|^p\right] \\ &\leq (2aL(k+1))^p \times \\ &\quad \times \frac{1}{k+1}\sum_{j=0}^{k}E\left[\|\theta_{t-j} - \theta_{t-j-1}\|^p\right],\end{aligned} \tag{40}$$

where (40) follows from Jensen's inequality. Combining (40) with Proposition 1 we arrive at

$$\left\|E\left[\left(h,\xi^N_{t-k,\theta_{t-k}}\right)|\mathcal{G}_{t-k}\right] - \left(h,\Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right)\right\|_p \leq 2aL(k+1)\frac{c_\kappa}{\sqrt{N}}, \tag{41}$$

where $c_\kappa < \infty$ is a constant independent of N, t and $\{\theta_n\}_{n\geq 0}$.

If we now insert (33) and (41) into (31) we obtain the relationship

$$\left\| \left(h, \xi^N_{t-k,\theta_{t-k}}\right) - \left(h, \Psi^{\theta_t}_{t-k}\left(\xi^N_{t-k-1,\theta_{t-k-1}}\right)\right) \right\|_p \leq \frac{c + 2aL(k+1)c_\kappa}{\sqrt{N}}, \tag{42}$$

where the numerator is finite and constant w.r.t. N, $\{\theta_n\}_{n\geq 0}$ and t. At this point, we only need to substitute the latter inequality backwards. Indeed, if we plug (42), with $h = \tilde{f}_{t-k}$, into (30) and then substitute the resulting bound, together with (29), into (28), we arrive at

$$\left\|(f, \xi^N_{t,\theta_t}) - (f, \xi_{t,\theta_t})\right\|_p \leq \frac{2\|f\|_\infty a^m \epsilon_\tau^{-1}}{\sqrt{N}} \sum_{k=0}^{t} \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{k}{m} \rfloor} (\bar{C}_0 + \bar{C}_1 k), \tag{43}$$

where $\bar{C}_0 = c + 2aLc_\kappa$ and $\bar{C}_1 = \tilde{C}_0 \vee 2aLc_\kappa$.

What remains to be proved is that the sum in (43) admits an upper bound $\bar{C} < \infty$ independent of t. To show this, we decompose

$$\sum_{k=0}^{t} \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{k}{m} \rfloor} (\bar{C}_0 + \bar{C}_1 k) = \bar{C}_0 \sum_{k=0}^{t} \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{k}{m} \rfloor} + \bar{C}_1 \sum_{k=0}^{t} k \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{k}{m} \rfloor} \tag{44}$$

and note that each term in (44) can be written as a sum of convergent series. Indeed, for the first term we have

$$\begin{aligned} \sum_{k=0}^{t} \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{k}{m} \rfloor} &\leq m \sum_{k=0}^{\infty} \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{k} && (45) \\ &= m a^{m-1} \epsilon_\tau^{-2}, && (46) \end{aligned}$$

where the inequality (45) is obtained from the identity $\sum_{k=0}^{\infty} r^{\lfloor \frac{k}{m} \rfloor} = m \sum_{k=0}^{\infty} r^k$ (for any $r \in (0,1)$) and (46) follows from the limit of the geometric series. For the second term in (44) we have

$$\begin{aligned} \sum_{k=0}^{t} k \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{k}{m} \rfloor} &\leq 2m \sum_{k=0}^{\infty} \left\lfloor \frac{k}{m} \right\rfloor \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{\lfloor \frac{k}{m} \rfloor} && (47) \\ &= 2m^2 \sum_{k=0}^{\infty} k \left(1 - \frac{\epsilon_\tau^2}{a^{m-1}}\right)^{k}, && (48) \\ &= 2m^2 \frac{1 - \epsilon_\tau^2 a^{-(m-1)}}{\epsilon_\tau^2 a^{-2(m-1)}}, && (49) \end{aligned}$$

where (47) follows from the inequality $k \leq 2m\lfloor \frac{k}{m} \rfloor$ (for $k = 0, 1, 2, ...$ and $m \geq 1$), (48) holds because of the identity $\sum_{k=0}^{\infty} \lfloor \frac{k}{m} \rfloor r^{\lfloor \frac{k}{m} \rfloor} = m \sum_{k=0}^{\infty} k r^k$ (for any $r \in (0,1)$) and (49) is readily obtained from the limit $\sum_{k=0}^{\infty} k r^k = \frac{r}{(1-r)^2}$ (for $|r| < 1$).

To conclude the proof, we simply put (43), (44), (46) and (49) together, to obtain the desired inequality (26) with

$$\bar{C} = 2\|f\|_\infty a^m \epsilon_\tau^{-1} \left(\bar{C}_0 m a^{m-1} \epsilon_\tau^{-2} + 2\bar{C}_1 m^2 \frac{1 - \epsilon_\tau^2 a^{-(m-1)}}{\epsilon_\tau^2 a^{-2(m-1)}}\right) \leq 4\|f\|_\infty (\bar{C}_0 \vee \bar{C}_1) \epsilon_\tau^{-3} a^{3m} \tag{50}$$

and $\bar{C}_0 \vee \bar{C}_1 \leq a(c + \tilde{C}_0 + 2Lc_\kappa)$. $\square$

5.3 Uniform convergence of the nested particle filter

Lemma 3 can be used to obtain bounds for the errors in the computation of the weights of Algorithm 2. Based on this result, it is possible to show that the overall procedure converges uniformly over time given the assumptions in Section 5.2, and provide an error rate. This is explicitly given by the following theorem.

Theorem 1 *Let $\{y_t\}_{t\geq 1}$ be an arbitrary sequence of observations, let D_θ be a compact set and select a jittering kernel $\kappa_{N,\mathsf{p}}$ from the family in Eq. (8). If assumptions A.1, A.3, A.4 and A.5 are satisfied, then*

$$\lim_{N\to\infty} \sup_{t\geq 0} \|(h,\mu_t^N) - (h,\mu_t)\|_p = 0$$

for any $h \in B(D_\theta)$ and $1 \leq p \leq \mathsf{p}$. If, additionally, the exponential stability assumption A.2 holds, then there exists $C < \infty$, independent of N and t, such that

$$\sup_{t\geq 0} \left\|(h,\mu_t^N) - (h,\mu_t)\right\|_p \leq N^{-\frac{1}{2}+\epsilon} + CN^{-\epsilon\frac{\bar{b}_2}{1+\log(a)}}$$

for any $0 < \epsilon < \frac{1}{2}$, where $C < \infty$ is a constant independent of N and t, while a and $\bar{b}_2$ are the constants specified in assumptions A.3 and A.2.

Proof: Choose some integer $T > 0$. We look into the error $\|(h,\mu_t^N) - (h,\mu_t)\|_p$ for $t < T$ and $t \geq T$ separately.

For any $t \geq T$, the difference $(h,\mu_t^N) - (h,\mu_t)$ can be decomposed as

$$\begin{aligned}(h,\mu_t^N) - (h,\mu_t) &= \sum_{k=0}^{T-1} \left(h,\Lambda_{t|t-k}(\mu_{t-k}^N)\right) - \left(h,\Lambda_{t|t-k-1}(\mu_{t-k-1}^N)\right) \\ &\quad + \left(h,\Lambda_{t|t-T}(\mu_{t-T}^N)\right) - \left(h,\Lambda_{t|t-T}(\mu_{t-T})\right). \end{aligned} \tag{51}$$

The last term on the right hand side of (51) can be bounded using A.1, namely

$$\left|\left(h,\Lambda_{t|t-T}(\mu_{t-T}^N)\right) - \left(h,\Lambda_{t|t-T}(\mu_{t-T})\right)\right| \leq \mathcal{S}(h,T), \tag{52}$$

where $\mathcal{S}(h,T)$ is independent of N and t, and $\lim_{T\to\infty} \mathcal{S}(h,T) = 0$ for every $h \in B(D_\theta)$. Minkowski's inequality, together with (51) and (52), readily yields an upper bound for the approximation error, namely

$$\|(h,\mu_t^N) - (h,\mu_t)\|_p \leq \sum_{k=0}^{T-1} \left\|\left(h,\Lambda_{t|t-k}(\mu_{t-k}^N)\right) - \left(h,\Lambda_{t|t-k-1}(\mu_{t-k-1}^N)\right)\right\|_p + \mathcal{S}(h,T), \tag{53}$$

and all we need to do is to calculate suitable bounds for the terms in the summation above.

It is not difficult to show (see Definition 4) that, for any $\alpha \in \mathcal{P}(D_\theta)$,

$$\left(h,\Lambda_{t|t-k}(\alpha)\right) = \frac{\left(h\prod_{j=0}^{k-1} u_{t-j},\alpha\right)}{\left(\prod_{j=0}^{k-1} u_{t-j},\alpha\right)}, \tag{54}$$

where $u_t(\theta) = (g_{t,\theta}^{y_t}, \xi_{t,\theta})$. From (54), the k-th term in the summation of (53) can be rewritten as

$$\left(h, \Lambda_{t|t-k}(\mu_{t-k}^N)\right) - \left(h, \Lambda_{t|t-k-1}(\mu_{t-k-1}^N)\right) = \frac{\left(h \prod_{j=0}^{k-1} u_{t-j}, \mu_{t-k}^N\right)}{\left(\prod_{j=0}^{k-1} u_{t-j}, \mu_{t-k}^N\right)} - \frac{\left(h \prod_{j=0}^{k-1} u_{t-j}, \Lambda_{t-k}(\mu_{t-k-1}^N)\right)}{\left(\prod_{j=0}^{k-1} u_{t-j}, \Lambda_{t-k}(\mu_{t-k-1}^N)\right)}$$

hence, by way of inequality (1), we obtain

$$\begin{aligned} \left\| \left(h, \Lambda_{t|t-k}(\mu_{t-k}^N)\right) - \left(h, \Lambda_{t|t-k-1}(\mu_{t-k-1}^N)\right) \right\|_p \quad \leq \\ a^k \left[\left\| \left(h \prod_{j=0}^{k-1} u_{t-j}, \mu_{t-k}^N \right) - \left(h \prod_{j=0}^{k-1} u_{t-j}, \Lambda_{t-k}(\mu_{t-k-1}^N) \right) \right\|_p + \right. \\ \left. \|h\|_\infty \left\| \left(\prod_{j=0}^{k-1} u_{t-j}, \mu_{t-k}^N \right) - \left(\prod_{j=0}^{k-1} u_{t-j}, \Lambda_{t-k}(\mu_{t-k-1}^N) \right) \right\|_p \right], \end{aligned} \tag{55}$$

where we have made use of assumption A.3 to obtain the factor a^k.

The two L_p norms on the right-hand side of (55) have the form $\|(v, \mu_n^N) - (v, \Lambda_n(\mu_{n-1}^N)\|_p$, for $n = t-k$ and $v \in B(D_\theta)$ (namely, $v = h \prod_{j=0}^{k-1} u_{t-j}$ in the first term and $v = \prod_{j=0}^{k-1} u_{t-j}$ in the second term). Therefore, we now seek a bound for $\|(v, \mu_n^N) - (v, \Lambda_n(\mu_{n-1}^N)\|_p$ that can be substituted back into (55).

Recall that Algorithm 2 succesively produces the approximate measures $\bar{\mu}_{n-1}^N = \frac{1}{N} \sum_{i=1}^N \delta_{\bar{\theta}_n^{(i)}}$, $\tilde{\mu}_n^N = \sum_{i=1}^N w_n^{(i)} \delta_{\bar{\theta}_n^{(i)}}$ and $\mu_n^N = \frac{1}{N} \sum_{i=1}^N \delta_{\theta_n^{(i)}}$. For the choice of kernel $\kappa_{N,\mathsf{p}}$ in (8) it is not difficult to show (see Appendix A) that

$$\|(v, \bar{\mu}_{n-1}^N) - (v, \mu_{n-1}^N)\|_p \leq \frac{s_1 \|v\|_\infty}{\sqrt{N}}, \tag{56}$$

where s_1 is a constant independent of n and N, and

$$\left\|(v, \mu_n^N) - (v, \tilde{\mu}_n^N)\right\|_p \leq \frac{s_2 \|v\|_\infty}{\sqrt{N}}, \tag{57}$$

where s_2 is also constant w.r.t. n and N (note that μ_n^N is obtained from $\tilde{\mu}_n^N$ by way of a multinomial resampling step). Therefore, if we use the triangle inequality

$$\begin{aligned} \|(v, \mu_n^N) - (v, \Lambda_n(\mu_{n-1}^N)\|_p \quad \leq \quad & \|(v, \mu_n^N) - (v, \tilde{\mu}_n^N)\|_p + \|(v, \tilde{\mu}_n^N) - (v, \Lambda_n(\bar{\mu}_{n-1}^N))\|_p \\ & + \|(v, \Lambda_n(\bar{\mu}_{n-1}^N)) - (v, \Lambda_n(\mu_{n-1}^N))\|_p \end{aligned} \tag{58}$$

and realise that, by way of (1) and assumption A.3,

$$\begin{aligned} \|(v, \Lambda_n(\bar{\mu}_{n-1}^N)) - (v, \Lambda_n(\mu_{n-1}^N))\|_p \quad = \quad & \left\| \frac{(vu_n, \bar{\mu}_{n-1}^N)}{(u_n, \bar{\mu}_{n-1}^N)} - \frac{(vu_n, \mu_{n-1}^N)}{(u_n, \mu_{n-1}^N)} \right\|_p \\ \leq \quad & a\|(vu_n, \bar{\mu}_{n-1}^N) - (u_n, \mu_{n-1}^N)\|_p + a\|v\|_\infty \|(u_n, \bar{\mu}_{n-1}^N) - (u_n, \mu_{n-1}^N)\|_p, \end{aligned} \tag{59}$$

then it is straightforward to take (59), (56) and (57) together and substitute them into (58) to obtain

$$\|(v, \mu_n^N) - (v, \Lambda_n(\mu_{n-1}^N)\|_p \leq \frac{\|v\|_\infty (2as_1 + s_2)}{\sqrt{N}} + \|(v, \tilde{\mu}_n^N) - (v, \Lambda_n(\bar{\mu}_{n-1}^N))\|_p \tag{60}$$

and only the second term on the right hand side of the inequality above remains to be bounded.

However, by the the construction of $\tilde{\mu}_n^N$ and Definition 4 (of Λ_n) we have

$$\begin{aligned}\|(v,\tilde{\mu}_n^N)-(v,\Lambda_n(\bar{\mu}_{n-1}^N))\|_p &= \left\|\frac{(vu_n^N,\bar{\mu}_{n-1}^N)}{(u_n^N,\bar{\mu}_{n-1}^N)}-\frac{(vu_n,\bar{\mu}_{n-1}^N)}{(u_n,\bar{\mu}_{n-1}^N)}\right\|_p \\ &\leq a\|(vu_n^N,\bar{\mu}_{n-1}^N)-(vu_n,\bar{\mu}_{n-1}^N)\|_p+\|v\|_\infty a\|(u_N^N,\bar{\mu}_{n-1}^N)-(u_n,\bar{\mu}_{n-1}^N)\|_p.\end{aligned} \tag{61}$$

Again, the two terms on the right hand side of the inequality (61) have essentially the same form, hence it is enough to analyse the first one. Writing the integrals w.r.t. $\bar{\mu}_{n-1}^N$ explicitly, extracting $v\leq\|v\|_\infty$ as a common factor and then applying Minkowski's inequality yields

$$\|(vu_n^N,\bar{\mu}_{n-1}^N)-(vu_n,\bar{\mu}_{n-1}^N)\|_p\leq\frac{\|v\|_\infty}{N}\sum_{i=1}^N\|u_n^N(\bar{\theta}_n^{(i)})-u_n(\bar{\theta}_n^{(i)}\|_p,$$

which, expanding the functions u_n^N and u_n as integrals w.r.t. $\xi^N_{n,\bar{\theta}_n^{(i)}}$ and $\xi_{n,\bar{\theta}_n^{(i)}}$, respectively, becomes

$$\|(vu_n^N,\bar{\mu}_{n-1}^N)-(vu_n,\bar{\mu}_{n-1}^N)\|\leq\frac{\|v\|_\infty}{N}\sum_{i=1}^N\left\|(g^{y_n}_{n,\bar{\theta}_n^{(i)}},\xi^N_{n,\bar{\theta}_n^{(i)}})-(g^{y_n}_{n,\bar{\theta}_n^{(i)}},\xi_{n,\bar{\theta}_n^{(i)}})\right\|_p. \tag{62}$$

However, by assumption A.3, $\sup_{n\geq0,\theta\in D_\theta,x\in\mathcal{X}}g^{y_n}_{n,\theta}\leq1$, hence (62) can be extended as

$$\|(vu_n^N,\bar{\mu}_{n-1}^N)-(vu_n,\bar{\mu}_{n-1}^N)\|_p\leq\frac{\|v\|_\infty}{N}\sum_{i=1}^N\sup_{\ell\in B(\mathcal{X}):\|\ell\|_\infty\leq1}\left(\sup_{n\geq0}\left\|(\ell,\xi^N_{n,\bar{\theta}_n^{(i)}})-(\ell,\xi_{n,\bar{\theta}_n^{(i)}})\right\|_p,\right) \tag{63}$$

where the terms $\sup_{n\geq0}\|(\ell,\xi^N_{n,\bar{\theta}_n^{(i)}})-(\ell,\xi_{n,\bar{\theta}_n^{(i)}})\|_p$ can be controlled by way of Lemma 3. To be specific, there exists a finite constant $\bar{C}$ independent of N and n such that

$$\sup_{n\geq0}\|(\ell,\xi^N_{n,\bar{\theta}_n^{(i)}})-(\ell,\xi_{n,\bar{\theta}_n^{(i)}})\|_p\leq\frac{\bar{C}}{\sqrt{N}}. \tag{64}$$

From (50) we readily see that there exists a constant $C^*<\infty$, independent of n, N, a and ℓ, such that $\bar{C}\leq C^*\|\ell\|_\infty a^{3m+1}$, hence

$$\sup_{\ell\in B(\mathcal{X}):\|\ell\|_\infty\leq1}\bar{C}\leq C^*a^{3m+1}<\infty. \tag{65}$$

Substituting (64) back into (63) and using (65) yields

$$\sup_{n\geq0}\|(vu_n^N,\bar{\mu}_{n-1}^N)-(vu_n,\bar{\mu}_{n-1}^N)\|_p\leq\frac{\|v\|_\infty C^*a^{3m+1}}{\sqrt{N}}. \tag{66}$$

From (66), we can substitute back into the sequence of inequalities that starts at (53). In particular, inserting (66) into (61) yields

$$\sup_{n\geq0}\|(v,\tilde{\mu}_n^N)-(v,\Lambda_n(\bar{\mu}_{n-1}^N))\|_p\leq\frac{2\|v\|_\infty C^*a^{3m+2}}{\sqrt{N}} \tag{67}$$

and plugging (67) into (60) we arrive at

$$\sup_{n\geq0}\|(v,\mu_n^N)-(v,\Lambda_n(\mu_{n-1}^N)\|_p\leq\frac{\|v\|_\infty\tilde{C}^*a^{3m+2}}{\sqrt{N}}, \tag{68}$$

where $\tilde{C}^* = 2C^* + 2s_1 + s_2$. The expression above yields bounds for the two terms on the right hand side of (55). Hence, substituting (68) into (55) we can write

$$\left\| \left(h, \Lambda_{t|t-k}(\mu_{t-k}^N)\right) - \left(h, \Lambda_{t|t-k-1}(\mu_{t-k-1}^N)\right) \right\|_p \quad \leq \quad \frac{2\|h\|_\infty \tilde{C}^* a^{3m+2+k}}{\sqrt{N}} \tag{69}$$

The inequality (69), in turn, provides bounds for each one of the terms in the summation of (53) which, taken together, lead to

$$\sup_{t \geq T} \|(h, \mu_t^N) - (h, \mu_t)\|_p \quad \leq \quad \frac{\|h\|_\infty \hat{C} T a^T}{\sqrt{N}} + \mathcal{S}(h, T), \tag{70}$$

where $\hat{C} = 2\tilde{C}^* a^{3m+2}$.

Next, we prove that a bound of the form in (70) also holds for $t < T$. In this case we can decompose the L_p norm of the approximation error as

$$\|(h, \mu_t^N) - (h, \mu_t)\|_p \quad \leq \quad \sum_{k=0}^{t-1} \|(h, \Lambda_{t|t-k}(\mu_{t-k}^N)) - (h, \Lambda_{t|t-k-1}(\mu_{t-k-1}^N))\|_p + \|(h, \Lambda_{t|0}(\mu_0^N) - (h, \Lambda_{t|0}(\mu_0))\|_p. \tag{71}$$

The sum on the right hand side of (71) has the same structure as the summation in (53), hence exactly the same argument leading to (70) (and bearing in mind that $t < T$) yields

$$\sup_{t<T} \sum_{k=0}^{t-1} \|(h, \Lambda_{t|t-k}(\mu_{t-k}^N)) - (h, \Lambda_{t|t-k-1}(\mu_{t-k-1}^N))\|_p \quad \leq \quad \frac{\|h\|_\infty \hat{C} T a^T}{\sqrt{N}} \tag{72}$$

which is the same bound as in (70) except for the residual $\mathcal{S}(h, T)$. As for the last term in (71), recall from (54) that $(h, \Lambda_{t|0}(\alpha)) = (h \prod_{j=0}^{t-1} u_{t-j}, \alpha)/(\prod_{j=0}^{t-1} u_{t-j}, \alpha)$ which, combined with (1), yields

$$\begin{aligned} \left\| (h, \Lambda_{t|0}(\mu_0^N)) - (h, \Lambda_{t|0}(\mu_0)) \right\|_p \quad \leq \quad & a^t \left[\left\| \left(h \prod_{j=0}^{t-1} u_{t-j}, \mu_0^N \right) - \left(h \prod_{j=0}^{t-1} u_{t-j}, \mu_0 \right) \right\|_p \right. \\ & \left. + \|h\|_\infty \left\| \left(\prod_{j=0}^{t-1} u_{t-j}, \mu_0^N \right) - \left(\prod_{j=0}^{t-1} u_{t-j}, \mu_0^N \right) \right\|_p \right]. \end{aligned}$$

Since μ_0^N is a random measure constructed with N i.i.d. samples from the distribution with measure μ_0, it is straightforward to show that there is a constant $\bar{c}_0 < \infty$, independent of N and t, such that

$$\left\| (h, \Lambda_{t|0}(\mu_0^N)) - (h, \Lambda_{t|0}(\mu_0)) \right\|_p \leq \frac{a^t \|h\|_\infty \bar{c}_0}{\sqrt{N}}. \tag{73}$$

If we recall that $t < T$ and put together (71), (72) and (73) then we readily obtain the bound

$$\sup_{t<T} \|(h, \mu_t^N) - (h, \mu_t)\|_p \leq \frac{\mathsf{C} \|h\|_\infty T a^T}{\sqrt{N}}, \tag{74}$$

where $\mathsf{C} = \hat{C} \vee \bar{c}_0$ is a finite constant independent of N, T and h.

Combining the inequalities (70) and (74) we have the error bound

$$\sup_{t\geq 0} \|(h,\mu_t^N) - (h,\mu_t)\|_p \leq \frac{\mathsf{C}\|h\|_\infty T a^T}{\sqrt{N}} + \mathcal{S}(h,T) \tag{75}$$

that holds for any positive integer $T < \infty$. In particular, we can choose $T = T_N^\epsilon$ such that $\mathsf{C}\|h\|_\infty T_N^\epsilon a^{T_N^\epsilon} \leq N^\epsilon$ for any $0 < \epsilon < \frac{1}{2}$. It is sufficient to set

$$T_N^\epsilon = \left\lfloor \frac{\epsilon \log(N) - \log(\mathsf{C}\|h\|_\infty)}{1 + \log(a)} \right\rfloor \tag{76}$$

in order to substitute $T = T_N^\epsilon$ in (75) and obtain

$$\sup_{t\geq 0} \|(h,\mu_t^N) - (h,\mu_t)\|_p \leq \frac{1}{N^{\frac{1}{2}-\epsilon}} + \mathcal{S}(h,T_N^\epsilon). \tag{77}$$

Since $\lim_{N\to\infty} T_N^\epsilon = \infty$ for every $\epsilon \in (0, \frac{1}{2})$, then assumption A.1 implies that

$$\lim_{N\to\infty} \mathcal{S}(h,T_N^\epsilon) = 0$$

and, as a consequence of (77), $\lim_{N\to\infty} \sup_{t\geq 0} \|(h,\mu_t^N) - (h,\mu_t)\|_p = 0$.

To complete the proof, we observe that assumption A.2 combined with (76) yields

$$\mathcal{S}(h,T_N^\epsilon) \leq C N^{-\epsilon \frac{\bar{b}_2}{1+\log(a)}}, \tag{78}$$

where $C = (\mathsf{C}\|h\|_\infty)^{\frac{\bar{b}_2}{1+\log(a)}} < \infty$ is independent of N and t. Combining (78) with (77) yields the explicit error bound in the statement of Theorem 1.

□

Remark 6 *While the convergence of Algorithm 2 can be guaranteed without assumption A.2, the latter is necessary in order to obtain the error bound in the statement of Theorem 1. To be specific, we need to specify how fast the error $\mathcal{S}(h,T)$ vanishes in order to compute an explicit error bound. This is given by assumption A.2, which describes a feature of the state-space model (rather than a feature of the algorithm).*

5.4 Parameter identification

The uniform convergence result of Theorem 1 implies that the vector of model parameters can be estimated exactly (as $t \to \infty$) provided that the sequence of observations is informative enough to guarantee that the posterior probability mass asymptotically concentrates around a single point in the parameter space D_θ. To be specific, in this section we assume that there exists $\theta_* \in D_\theta$ (which may be thought of as the "true value of Θ) such that

$$\lim_{t\to\infty} \mu_t = \delta_{\theta_*} \tag{79}$$

for the available sequence of observations $\{y_t\}_{t>0}$ and then proceed to show that $\mu_t^N \to \delta_{\theta_*}$ as $t \to \infty$, in a sense to be made precise. The existence of such θ_* is not a strong assumption. In [28] it is shown that, provided the parameter is "identifiable", meaning that

$$\theta_1 = \theta_2 \Leftrightarrow \lim_{t\to\infty} \phi_{t,\theta_1} = \lim_{t\to\infty} \phi_{t,\theta_2},$$

then the limit in (79) holds a.s. under mild assumptions.

Let $\Omega = \{h_i \in B(D_\theta) : \|h_i\|_\infty \leq 1, i \geq 1\}$ be a convergence determining set [2, Theorem 2.18] and define the distance $d_\Omega : \mathcal{P}(D_\theta) \times \mathcal{P}(D_\theta) \to [0, +\infty)$ as

$$d_\Omega(\alpha, \eta) \triangleq \sum_{i \geq 1} \frac{1}{2^i} |(h_i, \alpha) - (h_i, \eta)|$$

for any $\alpha, \eta \in \mathcal{P}(D_\theta)$. The existence of Ω is granted by [2, Theorem 2.18], while [2, Theorem 2.19] shows that a sequence of measures $\{\alpha_t \in \mathcal{P}(D_\theta)\}_{t \geq 1}$, converges weakly to another measure $\alpha \in \mathcal{P}(D_\theta)$ if, and only if, $\lim_{t\to\infty} d_\Omega(\alpha_t, \alpha) = 0$. The following result regarding the asymptotic identification of the system parameters is a fairly direct consequence of Theorem 1.

Theorem 2 *Let D_θ be a compact set and $\kappa_{N,\mathfrak{p}}$ a kernel of the class in Eq. (8). If assumptions A.1–A.5 hold and there exists $\theta_* \in D_\theta$ such that $\lim_{t\to\infty} \mu_t = \delta_{\theta_*}$, then, for any $0 < \epsilon < \frac{1}{2}$,*

$$\limsup_{t\to\infty} E\left[d_\Omega(\mu_t^N, \delta_{\theta_*})\right] \leq N^{-\frac{1}{2}+\epsilon} + CN^{-\epsilon\frac{\bar{b}_2}{1+\log(a)}} + 2^{-N+1} \tag{80}$$

where C, $\bar{b}_2$ and a are finite constants independent of N and t. In particular,

$$\lim_{N\to\infty} \limsup_{t\to\infty} E\left[d_\Omega(\mu_t^N, \delta_{\theta_*})\right] = 0.$$

Proof: We start with the triangle inequality

$$\sup_{n\geq t} E\left[d_\Omega(\mu_n^N, \delta_{\theta_*})\right] \leq \sup_{n\geq t} \left(E\left[d_\Omega(\mu_n^N, \mu_n)\right] + d_\Omega(\mu_n, \delta_{\theta_*})\right). \tag{81}$$

If we choose an integer $K \geq 1$ and expand d_Ω, the term $d_\Omega(\mu_n^N, \mu_n)$ can be upper bounded as

$$\begin{aligned} d_\Omega(\mu_n^N, \mu_n) &= \sum_{i=1}^{K} \frac{1}{2^i} |(h_i, \mu_n^N) - (h_i, \mu_n)| + \sum_{j>K} \frac{1}{2^j} |(h_i, \mu_n^N) - (h_i, \mu_n)| \\ &\leq \sum_{i=1}^{K} \frac{1}{2^i} |(h_i, \mu_n^N) - (h_i, \mu_n)| + \frac{1}{2^{K-1}}, \end{aligned} \tag{82}$$

where the inequality follows from bounding $|(h_i, \mu_n^N) - (h_i, \mu_n)| \leq 2$ and then computing $\sum_{j>K} 2^{-j} = 2^{-K}$. From (82), we readily obtain

$$\begin{aligned} \sup_{n\geq t} E\left[d_\Omega(\mu_n^N, \mu_n)\right] &\leq \sum_{i=1}^{K} \frac{1}{2^i} \sup_{n\geq t} \|(h_i, \mu_n^N) - (h_i, \mu_n)\|_1 + \frac{1}{2^{K-1}} \\ &\leq e(N)\left(1 - \frac{1}{2^{K-1}}\right) + \frac{1}{2^{K-1}}, \end{aligned} \tag{83}$$

where we have applied the identity $\sum_{i=1}^{K} 2^{-i} = 1 - 2^{-K+1}$ and the inequality

$$\sup_{n\geq t} \|(h_i, \mu_n^N) - (h_i, \mu_n)\|_1 \leq N^{-\frac{1}{2}+\epsilon} + CN^{-\epsilon\frac{\bar{b}_2}{1+\log(a)}} \triangleq e(N). \tag{84}$$

The latter follows from Theorem 1, with arbitrary $\epsilon \in (0, \frac{1}{2})$ and finite constants C, $\bar{b}_2$ and a independent of N and t. The inequality (83) is valid for any K, hence if we choose $N = K$ it readily follows that

$$\sup_{n \geq t} E\left[d_\Omega(\mu_n^N, \mu_n)\right] \leq e(N) + 2^{-N+1}. \tag{85}$$

If we now substitute (85) into (81) we obtain

$$\sup_{n \geq t} E\left[d_\Omega(\mu_n^N, \delta_{\theta_*})\right] \leq e(N) + 2^{-N+1} + \sup_{n \geq t} d_\Omega(\mu_n, \delta_{\theta_*})$$

and taking the limit as $t \to \infty$ yields

$$\limsup_{t \to \infty} E\left[d_\Omega(\mu_t^N, \delta_{\theta_*})\right] \leq e(N) + 2^{-N+1},$$

since $\lim_{t\to\infty} \mu_t = \delta_{\theta_*}$ by assumption. Finally, note that $e(N) + 2^{-N+1}$ is exactly the bound in (80).

□

6 Numerical results

6.1 Simulation setup

We present some computer simulation results to illustrate the numerical performance of the proposed nested particle filtering scheme (Algorithm 2) with long sequences of observations. A study numerical of convergence with increasing number of particles is presented in [10]. Let us consider a 3-dimensional Lorenz system [26] with additive dynamical noise and partial noisy observations [7]. The state of this system is a 3-dimensional stochastic process $\{X(s)\}_{s\in(0,\infty)}$, taking values on $\mathbb{R}^3$, which evolves over time according to the stochastic differential equations

$$dX_1 = -S(X_1 - Y_1)ds + dW_1, \quad dX_2 = (RX_1 - X_2 - X_1X_3)\,ds + dW_2, \quad dX_3 = (X_1X_2 - BX_3)\,ds + dW_3,$$

where $\{W_i(s)\}_{s\in(0,\infty)}$, $i = 1, 2, 3$, are independent 1-dimensional Wiener processes and $(S, R, B) \in \mathbb{R}$ are unknown model parameters. To put this system within the framework of this paper, we apply Euler's method with integration step $\Delta > 0$ to obtain the stochastic difference equations

$$X_{1,t} = X_{1,t-1} - \Delta S(X_{1,t-1} - X_{2,t-1}) + \sqrt{\Delta}U_{1,t}, \tag{86}$$

$$X_{2,t} = X_{2,t-1} + \Delta(RX_{1,t-1} - X_{2,t-1} - X_{1,t-1}X_{3,t-1}) + \sqrt{\Delta}U_{2,t}, \tag{87}$$

$$X_{3,t} = X_{3,t-1} + \Delta(X_{1,t-1}X_{2,t-1} - BX_{3,t-1}) + \sqrt{\Delta}U_{3,t}, \tag{88}$$

where $\{U_{i,t}\}_{t=0,1,\ldots}$, $i = 1, 2, 3$, are independent sequences of i.i.d. normal r.v.'s with 0 mean and variance 1. The system is partially observed every 40 discrete-time steps, and the observations have the form $\{Y_n = (Y_{1,n}, Y_{3,n})\}_{n=1,2,\ldots}$, where

$$Y_{1,n} = k_o X_{1,40n} + V_{1,n}, \quad Y_{3,n} = k_o X_{3,40n} + V_{3,n}, \tag{89}$$

$k_o > 0$ is an unknown scale parameter and $\{V_{i,n}\}_{n=1,2,...}$, $i = 1, 3$, are independent sequences of i.i.d. normal random variables with zero mean and variance $\sigma^2 = \frac{1}{10}$.

Let $X_t = (X_{1,t}, X_{2,t}, X_{3,t})$ be the state vector, let $Y_n = (Y_{1,n}, Y_{3,n})$ be the observation vector and let $\Theta = (S, R, B, k_o)$ be the static and unknown model parameters to be estimated. It is simple to obtain the family of kernels $\tau_{t,\theta}(dx|x_{t-1})$ from Eqs. (86)–(88) and the likelihood $g^{y_n}_{n,\theta}(x_n)$ from Eq. (89). The sequences X_t and Y_n are defined on different time scales, however it is straightforward to construct a sequence $\hat{X}_n$, with the same time index as the observations, if we simply define $\hat{X}_n = X_{40n}$. The transition kernel for $\hat{X}_n$ is obtained by composing the kernels for X_t. In particular, for the purpose of implementing Algorithm 2, one can draw a sample $\hat{X}_n = \hat{x}_n$ conditional on θ and $\hat{X}_{n-1} = \hat{x}_{n-1}$, by successively simulating

$$\tilde{x}_t \sim \tau_{t,\theta}(dx|\tilde{x}_{t-1}), \quad t = 40(n-1)+1, ..., 40n,$$

where $\tilde{x}_{40(n-1)} = \hat{x}_{n-1}$ and $\hat{x}_n = \tilde{x}_{40n}$. The prior measure for the state variables is normal and independent of Θ, namely

$$X_0 \sim \mathcal{N}(x_*, v_0^2 \mathcal{I}_3),$$

where $x_* = (-5.91652; -5.52332; 24.5723)$ is the mean and $v_0^2\mathcal{I}_3$ is the covariance matrix, with $v_0^2 = 10$ and $\mathcal{I}_3$ the 3-dimensional identity matrix. The value x_* has been taken from a simulated trajectory of the deterministic Lorenz 63 model. In this way we ensure that the simulation for the stochastic model starts at a "reasonable" point in the state space.

The goal is to track the posterior probability measures of the parameters, $\mu_n(d\theta) = \mathbb{P}\{\Theta \in d\theta | Y_{1:n}\}$, $n = 1, 2, ...$, using Algorithm 2. We assume that the parameters are a priori independent, namely

$$S \sim \mathcal{U}(5, 20), \quad R \sim \mathcal{U}(18, 50), \quad B \sim \mathcal{U}(1, 8) \quad \text{and} \quad k_o \in \mathcal{U}(0.5, 3),$$

where $\mathcal{U}(a, b)$ is the uniform probability distribution in the interval (a, b). Therefore the prior measure μ_0 is uniform, with support $D_\theta = [5, 20] \times [18, 50] \times [18, 50] \times [1, 8] \times [0.5, 3]$.

In order to run Algorithm 2 we need to choose the number of particles in the state space, N, the number of particles in the parameter space M, and the jittering kernel $\kappa_{N,\mathsf{p}}$. For the set of computer experiments here, we have set $N = M = 300$ and the jittering kernel is selected as in (8), in particular

$$\kappa^{\theta_{n-1}}_{N,\mathsf{p}}(d\theta) = (1 - \epsilon_N)\delta_{\theta_{n-1}}(d\theta) + \epsilon_N \bar{\kappa}^{\theta_{n-1}}(\theta)d\theta,$$

where $\epsilon_N = \frac{1}{\sqrt{N}}$ and $\bar{\kappa}^{\theta_{n-1}}(\theta)$ is a truncated-Gaussian pdf with support D_θ and independent of N, namely

$$\bar{\kappa}^{\theta_{n-1}}(\theta) = \mathsf{c}_{n-1} \exp\left\{-\frac{1}{2}\left(\theta - \theta_{n-1}\right)^\top \mathcal{C}^{-1}\left(\theta - \theta_{n-1}\right)\right\}, \quad \theta \in D_\theta,$$

where the proportionality constant c_{n-1} is a function of θ_{n-1} and the (fixed) covariance matrix is

$$\mathcal{C} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 \\ 0 & 0 & 0 & \frac{1}{20} \end{bmatrix}.$$

6.2 Results

The actual parameter values used for the computer experiments in this section are $(S, R, B, k_o) = (10, 28, \frac{8}{3}, 0.8)$, which yield an underlying chaotic dynamics.

Figure 1 shows the posterior mean estimates of the parameters S, R, B and k_o obtained for a single simulation with $N = M = 300$ particles and a length of 1,000 continuous time units. Since the Euler's integration step is $\Delta = 10^{-3}$ continuous time units and observations are taken every 40Δ continuous time units, the simulation involves 10^6 discrete time steps and 25×10^3 observations vectors. At discrete time n, the posterior mean of the parameter vector $\Theta = (S, R, B, k_o)$ is computed as $\hat{\theta}_n^N = \frac{1}{N}\sum_{i=1}^N w_n^{(i)} \bar{\theta}_n^{(i)}$. In the same figure it can be seen that, after a relatively short convergence period, the estimates remain locked to the true parameter values (plotted with black solid lines). The posterior-mean approximation $\hat{\theta}_n^N$ is random and it only converges to the exact posterior mean as $N \to \infty$, hence some fluctuations can be observed over time. However, the amplitude of the fluctuations remains bounded and stable over the whole simulation run.

Figure 2 shows the normalised posterior standard deviation (NSTD) of the parameter estimates for the same simulation run. At each time n, this is computed for the j-th parameter, $j = 1, ..., 4$, as

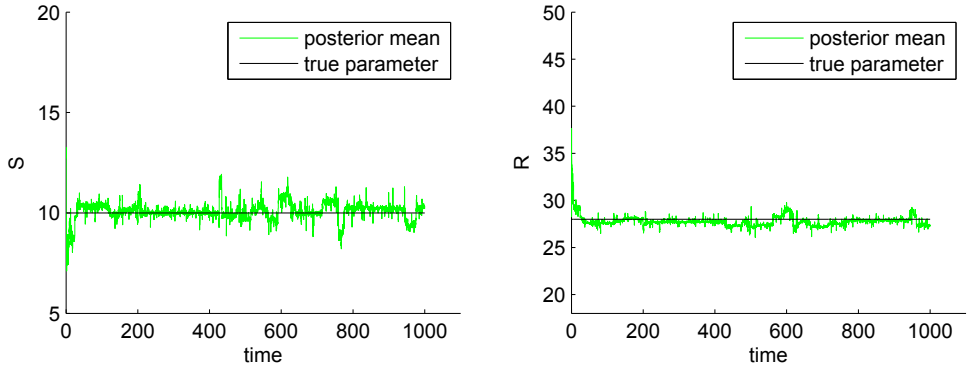
$$NSTD_{j,n} = \frac{\sqrt{\sum_{i=1}^N w_n^{(i)} (\bar{\theta}_{j,n}^{(i)} - \hat{\theta}_{j,n}^N)^2}}{\theta_j^*},$$

where θ_j^* is the true value of the j-th parameter (namely, $\theta_1^* = S = 10, \theta_2^* = R = 28, \theta_3^* = B = \frac{8}{3}$ and $\theta_4^* = k_o = 0.8$). Again, the NSTD is a random statistic and it displays fluctuations, however it can be seen that their amplitudes remain bounded and there is no apparent increase over time.

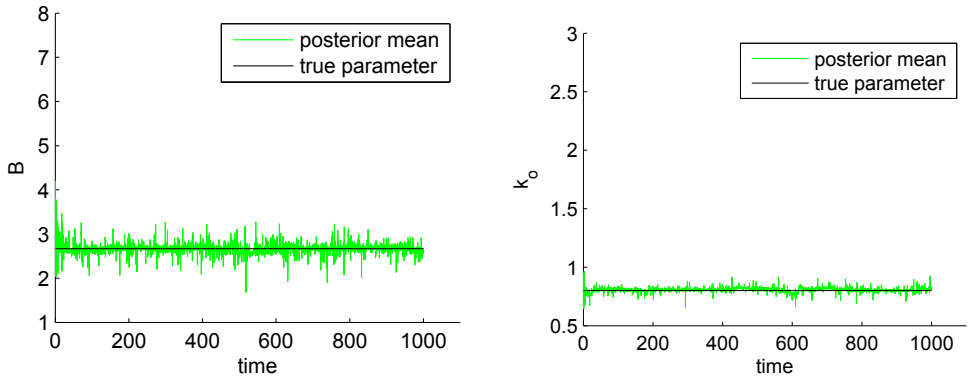
Figure 3 displays the errors between the posterior-mean estimates of the state variables and the actual values, for the same simulation run as in Figures 1 and 2. At discrete time n, the estimates are computed as $\hat{x}_{\ell,n}^N = \frac{1}{N}\sum_{i=1}^N w_n^{(i)} \sum_{j=1}^N \hat{x}_n^{(i,j)}$, for $\ell = 1, 2, 3$, and the errors displayed are of the form $e_{\ell,n}^N = \hat{x}_{\ell,n}^N - \hat{x}_n$. It can be seen that the errors are large at the beginning of the simulation. This is a consequence of the initial uncertainty in the values of the fixed parameters. Once the parameter estimates have converged, the errors decrease substatially and remain bounded, stable and centred around 0 for the rest of the simulation.

Finally, we have carried out a set of 50 independent simulations in order to approximate the mean absolute error of the parameter (posterior-mean) estimates. For each simulation we have run the stochastic Lorenz 63 model for 400 continuous time units, which amounts to 400×10^3 discrete time steps and a sequence of $10,000$ observations. For each simulation and each time step, we have computed the absolute error of the posterior-mean estimate of each parameter. Then, we have averaged these errors over the 50 independent simulation runs.

Figure 4 displays the mean absolute error for each parameter, S, R, B and k_o, over time. We observe that there is a convergence period and, after approximately 100 continuous time units, the error converges

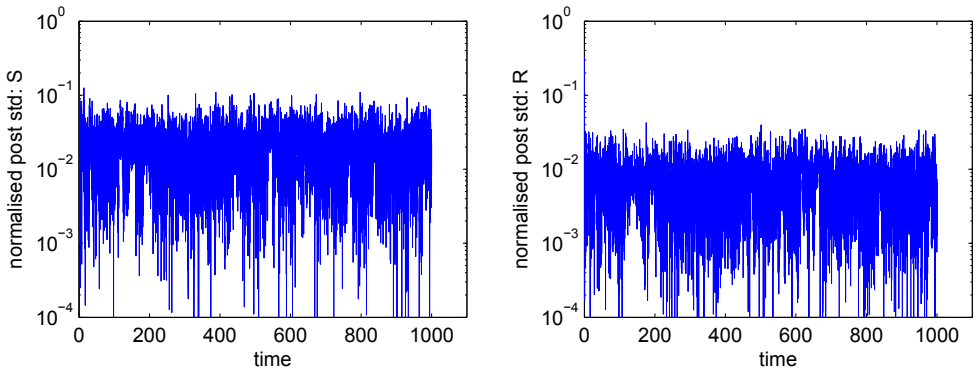


(a) Posterior-mean estimates of parameter S. (b) Posterior-mean estimates of parameter R.



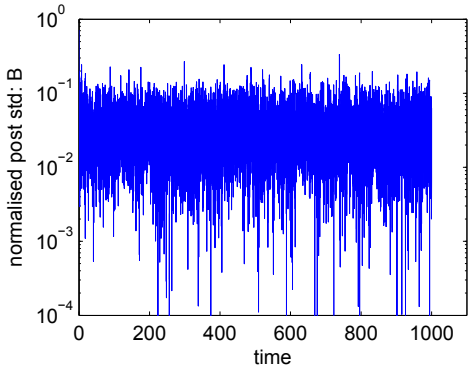
(c) Posterior-mean estimates of parameter B. (d) Posterior-mean estimates of parameter k_o.

Figure 1: Evolution of the posterior-mean estimates of the Lorenz 63 model parameters S, R, B and k_o over time. The horizontal axes are labeled with continuous time units. After Euler's discretisation, each continuous time unit amounts to 1,000 discrete time steps (hence, 1 million time steps for the complete simulation), with one observation vector every 40 discrete-time steps. The number of particles is $N = M = 300$. The vertical axes extend over the exact prior support for each parameter, i.e., $S \in [5, 20], R \in [18, 50], B \in [1, 8]$ and $k_o \in [0.5, 3]$.

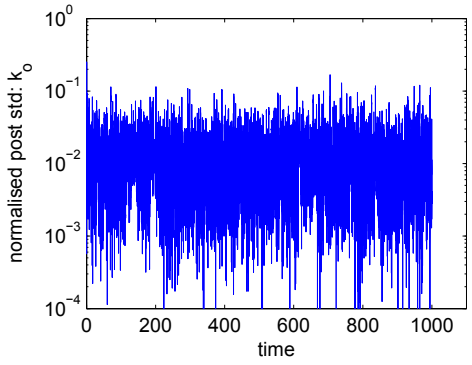


(a) Normalised posterior standard deviation of the parameter S.

(b) Normalised posterior standard deviation of the parameter R.

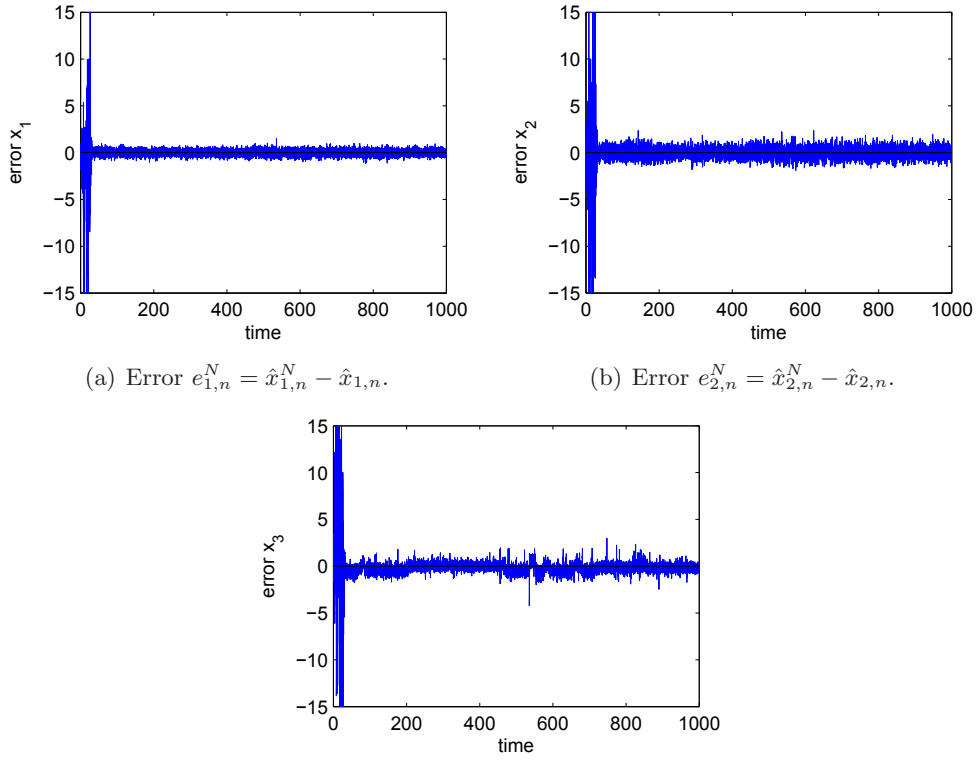


(c) Normalised posterior standard deviation of the parameter B.



(d) Normalised posterior standard deviation of the parameter k_o.

Figure 2: Evolution of the normalised posterior standard deviation of the Lorenz 63 model parameters S, R, B and k_o over time. The horizontal axes are labeled with continuous time units. After Euler's discretisation, each continuous time unit amounts to 1,000 discrete time steps, with one observation vector every 40 discrete-time steps. The number of particles is $N = M = 300$.

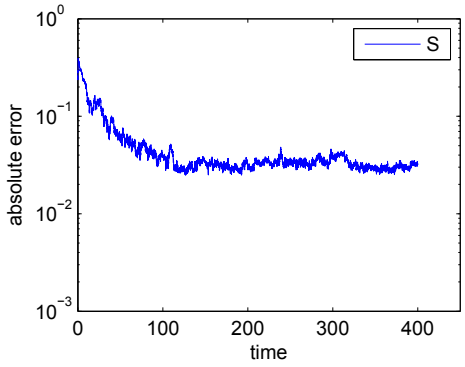


(a) Error $e^N_{1,n} = \hat{x}^N_{1,n} - \hat{x}_{1,n}$.

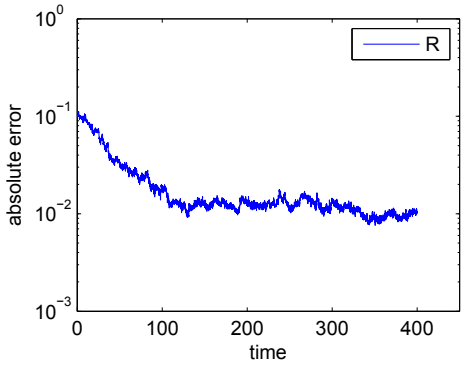
(b) Error $e^N_{2,n} = \hat{x}^N_{2,n} - \hat{x}_{2,n}$.

(c) Error $e^N_{3,n} = \hat{x}^N_{3,n} - \hat{x}_{3,n}$.

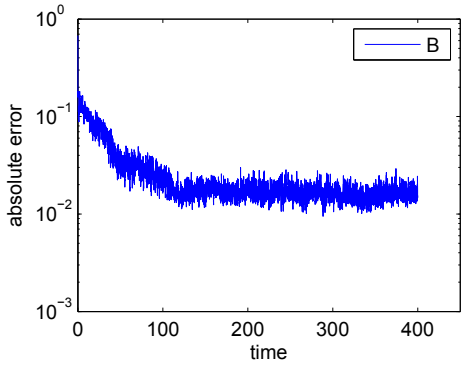
Figure 3: Evolution of the errors $e^N_{\ell,n} = \hat{x}^N_{\ell,n} - \hat{x}_n$, $\ell = 1, 2, 3$, for the state variables of the Lorenz 63 model, where the estimates $\hat{x}^N_{\ell,n}$ are posterior means. The horizontal axes are labeled with continuous time units. After Euler's discretisation, each continuous time unit amounts to 1,000 discrete time steps (hence, 1 million time steps for the complete simulation), with one observation vector every 40 discrete-time steps. The number of particles is $N = M = 300$.



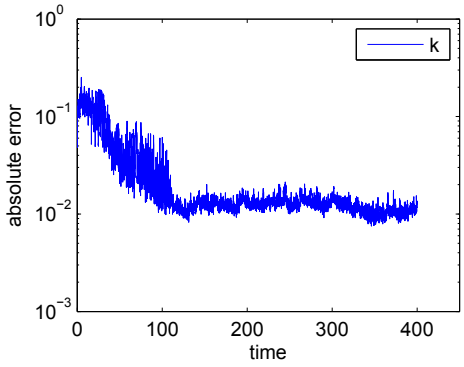
(a) Average absolute error of the posterior-mean estimates; parameter S.



(b) Average absolute error of the posterior-mean estimates; parameter R.



(c) Average absolute error of the posterior-mean estimates; parameter B.



(d) Average absolute error of the posterior-mean estimates; parameter k_o.

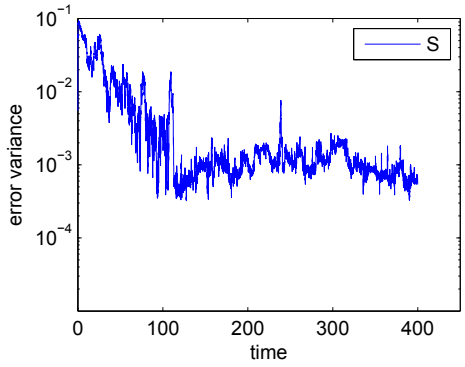
Figure 4: Absolute errors of the posterior-mean estimates of the Lorenz 63 model parameters, S, R, B and k_o, versus continuous time. The errors have been averaged over 50 independent simulation runs. The length of each simulation is 400 continuous time units, which amounts to 400×10^3 discrete-time steps after discretisation of the Lorenz 63 model, with a sequence of 10,000 observations.

to a steady value and remains stable for the rest of the simulation. The same kind of performance is observed for the variance of the absolute errors, computed over the same set of 50 independent simulations, and shown in Figure 5.

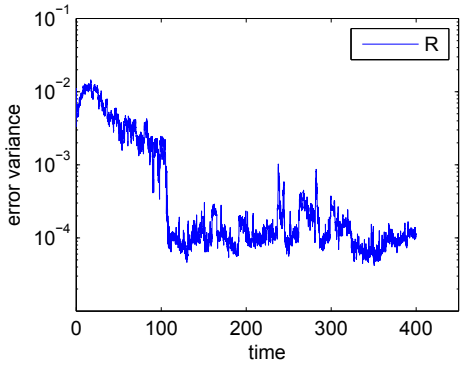
7 Conclusions

We have analysed the asymptotic convergence of a recursive Monte Carlo scheme, consisting of two (nested) layers of particle filters, for the approximation and tracking of the posterior probability distribution of the unknown parameters of a state-space Markov system. The algorithm is similar to the recently proposed SMC^2 method, however the scheme in this paper is purely recursive and, thus, potentially more useful for online implementations.

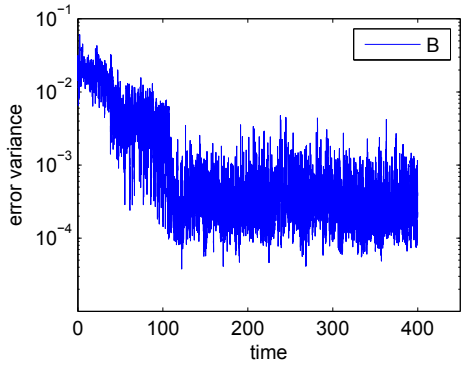
The theoretical contribution of the paper includes the analysis of the errors in the approximation of



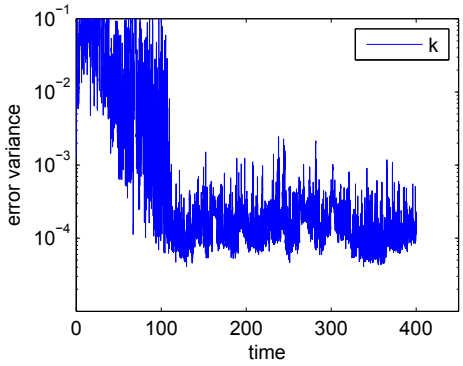
(a) Variance of the absolute error of the posterior-mean estimates; parameter S.



(b) Variance of the absolute error of the posterior-mean estimates; parameter R.



(c) Variance of the absolute error of the posterior-mean estimates; parameter B.



(d) Variance of the absolute error of the posterior-mean estimates; parameter k_o.

Figure 5: Variance of the absolute errors of the posterior-mean estimates of the Lorenz 63 model parameters, S, R, B and k_o, versus continuous time. The variances have been estimated from 50 independent simulation runs. The length of each simulation is 400 continuous time units, which amounts to 400×10^3 discrete-time steps after discretisation of the Lorenz 63 model, with a sequence of 10,000 observations.

integrals of bounded functions w.r.t. the posterior probability measure of the parameters. The analysis is carried out under regularity assumptions that include:

- The compactness of the parameter space.
- The stability of the sequence of posterior probability measures of the unknown parameters, $\{\mu_t\}$, w.r.t. the initial measure μ_0.
- A state space model that consists of a mixing Markov kernel and a normalised likelihood function with a positive lower bound. These regularity conditions are assumed to be satisfied uniformly over the parameter support. If this this assumption is met, then the classical results in [11] imply that the standard particle filters for the state space model of interest converge uniformly over time for any choice of the parameters in the support set D_θ.
- The Markov kernel has a pdf (w.r.t. the Lebesgue measure) which is Lipschitz continuous w.r.t. the vector of unknown parameters. The likelihood function in the model is also assumed to be Lipschitz continuous w.r.t. the parameters.

These assumptions are restrictive, yet they simply describe a model for which the standard particle filter would converge uniformly over time (were the parameters known) *and* for which small perturbations to the parameters yield small perturbations in the sequence of posterior probability measures (for the same sequence of observations). The convergence of the proposed recursive algorithm cannot be guaranteed if any of the assumptions above is not met (e.g., for models in which some specific choice of the parameters may yield an unstable behaviour).

The uniform convergence result in Theorem 1 has additional implications. In this paper, we have proved that, for a class of non-ambiguous models [28], the parameters can be identified, i.e., they can be estimated in an asymptotically exact manner (meaning that the sequence of approximate posterior measures generated by the algorithm converge to a delta measure).

Acknowledgements

The work of J. Míguez was partially supported by *Ministerio de Economía y Competitividad* of Spain (project TEC2012-38883-C02-01 COMPREHENSION) and the Office of Naval Research Global (award no. N62909-15-1-2011). Part of this work was carried out while J. M. was a visitor at the Department of Mathematics of Imperial College London, with partial support from an EPSRC Mathematics Platform grant. D. C. and J. M. would also like to acknowledge the support of the Isaac Newton Institute through the program "Monte Carlo Inference for High-Dimensional Statistical Models".

A A proof for inequality (56)

We need to prove that $\|(v,\bar{\mu}_{n-1}^N)-(v,\mu_{n-1}^N)\|_p \leq \frac{s_1\|v\|_\infty}{\sqrt{N}}$ for some $s_1 < \infty$ independent of N and $v \in B(D_\theta)$.

Recall that we draw the particles $\bar{\theta}_n^{(i)}$, $i=1,\ldots,N$, independently from the kernels $\kappa_{N,\mathsf{p}}^{\theta_{n-1}^{(i)}}$, $i=1,\ldots,N$, respectively, and start from the triangle inequality

$$\|(v,\bar{\mu}_{n-1}^N)-(v,\mu_{n-1}^N)\|_p \leq \|(v,\bar{\mu}_{n-1}^N)-(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)\|_p + \|(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)-(v,\mu_{n-1}^N)\|_p \tag{90}$$

where

$$(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N) = \frac{1}{N}\sum_{i=1}^N \int v(\theta)\kappa_{N,\mathsf{p}}^{\theta_{n-1}^{(i)}}(d\theta),$$

and then analyse the two terms on the right hand side of (90) separately.

Let $\mathcal{G}_{n-1}$ be the σ-algebra generated by the random particles $\{\bar{\theta}_{1:n-1}^{(i)},\theta_{0:n-1}^{(i)}\}_{1\leq i\leq N}$. Then

$$E\left[(v,\bar{\mu}_{n-1}^N)|\mathcal{G}_{n-1}\right] = \frac{1}{N}\sum_{i=1}^N \int v(\theta)\kappa_{N,\mathsf{p}}^{\theta_{n-1}^{(i)}}(d\theta) = (v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)$$

and the difference $(v,\bar{\mu}_{n-1}^N)-(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)$ can be written as

$$(v,\bar{\mu}_{n-1}^N)-(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N) = \frac{1}{N}\sum_{i=1}^N \bar{Z}_{n-1}^{(i)},$$

where the random variables $\bar{Z}_{n-1}^{(i)} = v(\bar{\theta}_n^{(i)}) - E[v(\bar{\theta}_n^{(i)})|\mathcal{G}_{n-1}]$, $i=1,...,N$, are conditionally independent (given $\mathcal{G}_{n-1}$), have zero mean and can be bounded as $|\bar{Z}_{n-1}^{(i)}| \leq 2\|v\|_\infty$. As a consequence, it is an exercise in combinatorics to show that

$$E\left[\left|(v,\bar{\mu}_{n-1}^N)-(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)\right|^p|\mathcal{G}_{n-1}\right] = E\left[\left|\frac{1}{N}\sum_{i=1}^N \bar{Z}_{n-1}^{(i)}\right|^p|\mathcal{G}_{n-1}\right] \leq \frac{\tilde{c}_1^p\|v\|_\infty^p}{N^{\frac{p}{2}}}, \tag{91}$$

where $\tilde{c}_1$ is a constant independent of N, n and v (actually, independent of the distribution of the $\bar{Z}_{n-1}^{(i)}$'s). From (91) we readily obtain that

$$\|(v,\bar{\mu}_{n-1}^N)-(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)\|_p \leq \frac{\tilde{c}_1\|v\|_\infty}{\sqrt{N}}. \tag{92}$$

For the remaining term in (90), namely, $\|(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)-(v,\mu_{n-1}^N)\|_p$, we simply note that

$$\begin{aligned}\left|(v,\kappa_{N,\mathsf{p}}\mu_{n-1}^N)-(v,\mu_{n-1}^N)\right| &= \left|\frac{1}{N}\sum_{i=1}^N\int\left(v(\theta)-v(\theta_{n-1}^{(i)})\right)\kappa_{N,\mathsf{p}}^{\theta_{n-1}^{(i)}}(d\theta)\right| \\ &\leq \frac{1}{N}\sum_{i=1}^N\int\left|v(\theta)-v(\theta_{n-1}^{(i)})\right|\kappa_{N,\mathsf{p}}^{\theta_{n-1}^{(i)}}(d\theta) \leq \frac{2\|v\|_\infty}{\sqrt{N}},\end{aligned} \tag{93}$$

where the last inequality follows from Proposition 1.

Substituting the inequalities (92) and (93) into Eq. (90) yields the desired conclusion, viz., Eq. (56), with constant $s_1 = 2+\tilde{c}_1$ independent of N.

References

[1] C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadić. Particle methods for change detection, system identification and control. *Proceedings of the IEEE*, 92(3):423–438, March 2004.

[2] A. Bain and D. Crisan. *Fundamentals of Stochastic Filtering*. Springer, 2008.

[3] O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[4] C. M. Carvalho, M. S. Johannes, H. F. Lopes, and N. G. Polson. Particle learning and smoothing. *Statistical Science*, 25(1):88–106, 2010.

[5] R. Chen and J. S. Liu. Mixture Kalman filters. *Journal of the Royal Statistics Society B*, 62:493–508, 2000.

[6] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2012.

[7] A. J. Chorin and P. Krause. Dimensional reduction for a Bayesian filter. *PNAS*, 101(42):15013–15017, October 2004.

[8] D. Crisan. Particle filters - a theoretical perspective. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 2, pages 17–42. Springer, 2001.

[9] D. Crisan and J. Miguez. Particle-kernel estimation of the filter density in state-space models. *Bernoulli*, 20(4):1879–1929, 2014.

[10] D. Crisan and J. Miguez. Nested particle filters for online parameter estimation in discrete-time state-space markov models. *arXiv*, 1308.1883v3 [stat.CO], 2015.

[11] P. Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[12] P. Del Moral and A. Guionnet. On the stability of interacting processes with applications to filtering and genetic algorithms. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 37(2):155–194, 2001.

[13] A. Doucet, N. de Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 1, pages 4–14. Springer, 2001.

[14] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York (USA), 2001.

[15] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo Sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[16] N. Gordon, D. Salmond, and A. F. M. Smith. Novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113, 1993.

[17] N. Kantas, A. Doucet, S. S. Singh, and J. M. Maciejowski. An overview of sequential monte carlo methods for parameter estimation in general state-space models. In *15th IFAC Symposium on System Identification*, volume 15, 2009.

[18] N. Kantas, A. Doucet, S. S. Singh, J. M. Maciejowski, and N. Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30:328–351, August 2015.

[19] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state-space models. *J. Comput. Graph. Statist.*, 1:1–25, 1996.

[20] G. Kitagawa. A self-organizing state-space model. *Journal of the American Statistical Association*, pages 1203–1215, 1998.

[21] H. R. Künsch. Recursive Monte Carlo filters: Algorithms and theoretical bounds. *The Annals of Statistics*, 33(5):1983–2021, 2005.

[22] H. R. Künsch. Particle filters. *Bernoulli*, 19(4):1391–1403, 2013.

[23] F. LeGland and L. Mevel. Recursive estimation in hidden Markov models. In *Proceedings of the 36th IEEE Conference on Decision and Control, 1997*, volume 4, pages 3468–3473. IEEE, 1997.

[24] J. Liu and M. West. Combined parameter and state estimation in simulation-based filtering. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 10, pages 197–223. Springer, 2001.

[25] J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, September 1998.

[26] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2):130–141, 1963.

[27] B. N. Oreshkin and M. J. Coates. Analysis of error propagation in particle filters with approximation. *The Annals of Applied Probability*, 21(6):2343–2378, 2011.

[28] A. Papavasiliou. Parameter estimation and asymptotic stability in stochastic filtering. *Stochastic Processes and Their Applications*, 116:1048–1065, 2006.

[29] M. K. Pitt and N. Shephard. Auxiliary variable based particle filters. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, chapter 13, pages 273–293. Springer, 2001.

[30] G. Poyiadjis, A. Doucet, and S. S. Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80, 2011.

[31] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, Boston, 2004.

[32] G. Storvik. Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions Signal Processing*, 50(2):281–289, February 2002.